THE UNIVERSITY
of EDINBURGH

IDENTIFICATION AND VALIDATION OF MUTATED SIGNALLING PATHWAYS IN CANCER

ALI ALSAADI

Thesis Submitted to The University of Edinburgh for the degree of Doctor of Philosophy
2016

Declaration

I declare that this thesis has been completely written and composed by myself along with the work detailed herein were entirely my own unless otherwise clearly acknowledged. This piece of work has been submitted for the degree of Doctor of Philosophy and has not been submitted for any other degree or professional qualifications.

Ali Alsaadi

Abstract

Genome sequencing is emerging as a powerful tool to identify the molecular mechanism of cancer progression. However, the software tools to define genomic and post-genomic mutations are just in its infancy. We have used a novel software algorithm to analyse the cancer genome by DNAseq and expressed cancer genome arising from transcription by RNAseq to define dominant sources of potentially expressed tumour-specific mutations and oncogenic targets. We focus primarily on the rare human pleomorphic sarcoma as a disease of high unmet clinical need but use a range of cancer models to accelerate the development of the pipeline.

First, we applied next generation sequencing of whole exomes of tumour tissues and two matched normal tissues (blood and "normal" tumour adjacent tissue) from a small set of patients to define parameters for use of the new software. The approaches identified significant mutations in tumour relative to germline DNA, but also in normal adjacent tissue, relative to normal germline, consistent with known field cancerization. Thus, in setting up the larger sequencing screen in the subsequent set of twenty cancer pleomorphic sarcoma cancer patients, whole exome sequencing was performed on tumour tissue and their matched normal adjacent tissues, rather than germline blood derived DNA, to define truly tumour-specific mutations. This approach provided sets of recurrent non-synonymous mutations in tumour tissue such as a transmembrane protease and suggests potential therapeutic targets for future focus that are highly tumour specific in pleomorphic sarcoma.

A major problem with using DNA genomics only to define drugable landscapes in cancer is that the tumour genome is static and the mutations do not reflect the expressed cancer landscape at the time of surgery. Thus, in a smaller subset of patients we also applied shotgun RNAseq to determine the number of expressed mutated genes. We defined within the parameters chosen, from 8-17% of the mutated genome is expressed as defined at the RNA level. However, to our surprise, there were an order of magnitude more RNA mutations that were not DNA encoded suggestive of RNA editing events. Each patient showed elevated RNA edits that were independent of each other suggesting a highly-patient, cancer-specific perturbation in the

specificity of the RNA editing machinery. We thus developed a cancer cell model to validate the RNA-editing software and we found we could recapitulate some of the RNA edits observed in clinical tumour tissue, in particular the signalling kinase in the MAP kinase-kinase-kinase-kinase super-family. It was interesting that RNA edits can often cluster in exon-intron boundaries suggesting a link to splicing and allows us to begin to produce "rules" for RNA editing. These data provide future direction to understand the role of RNA editing, as well as DNA encoded mutations, as mutagenic events and possible drugable targets in cancer signalling.

Lastly, novel or orphan mutant proteins observed in human cancers, whether from DNA encoded mutant proteins or from RNA-edited driven mutant protein synthesis require new tools and technologies to discover new oncogenic signalling mechanisms. We developed an SBP-tagged affinity purification method in combination with label-free SWATH mass spectrometry to identify a novel binding protein for the gain-of-function mutant protein in a key metastatic gene, ELMO1. This identified an elevated interaction with another oncogenic protein encoded by AGR2 gene and validates this proteomics discovery platform to further advance function of new mutated proteins.

In conclusion, we have applied and validated newly emerging software to begin to interrogate cancer tissue from patients of unmet clinical need in order to define new mechanisms of cancer progression and to define possibly new or better drug targets for new therapies. The data identified highly recurrent genome encoded mutations in human pleomorphic sarcoma and a potentially novel, targetable landscape represented by RNA editing driven mutant protein production. This will provide a foundation for future work on making better choices to advance our ability to improve patient management in human pleomorphic sarcoma.

Acknowledgement

This thesis would not have been possible without the help of my supervisor Professor Ted Hupp who continuously helped and supported me throughout my PhD. I am grateful to him for all of his advices and efforts he offered me during the past few years I spent in his laboratory.

I would like to thank Prof. Kathryn Ball for her support and ideas and all members of Hupp and Ball laboratory for their help and support.

A special thanks to my wife Ruqeya for her never ending support and encouragement towards completing my PhD. Thanks for being there through all the good and bad science days.

I would like to thank also my daughters Afra, Aisha and Hessa for lovely days we spent in Scotland.

An important thanks to my mother who always supports me and makes sure everything is fine with me, and thanks to all my family members.

I am grateful also to my country, United Arab Emirates, for their sponsorship of my PhD studies.

Table of Contents

Abbreviations

A3B	APOBEC3B
ACT	adoptive cellular therapy
AGR2	anterior gradient protein 2
AML	acute myeloid leukaemia
cDNA	complementary DNA
CES	collision energy spread
CGH	comparative genome hybridization
CML	chronic myelogenous leukaemia
CNVs	copy number variations
CTLA-4	cytotoxic T lymphocyte-associated protein 4
DCD	Dermcidin
DH	DbI homology
DHR	DOCK-homology regions
DOCK2	dedicator of cytokinesis 2
EAD	ELMO1 autoregulatory domain
EGFR	epidermal growth factor receptor
EID	ELMO1 inhibitory domain

ELMO1	engulfment and cell motility 1
ER	endoplasmic reticulum
FISH	fluorescence in situ hybridization
GEFs	guanine nucleotide exchange factors
GERD	gastroesophageal reflux disease
GIST	Gastrointestinal stromal tumour
HBV	hepatitis B virus
HCC	hepatocellular carcinoma
HGD	high-grade dysplasia
HLA	human leukocyte antigen
HNC	Head and neck cancer
HNSCC	Head and neck squamous cell carcinoma
HPV	human papilloma virus
ICGC	International Cancer Genome Consortium
IH C	immunohistochemistry
INDELS	insertions or deletions
MHC	major histocompatibility complex
MS	mass spectrometry

NGS	next generation sequencing
NMD	Nonsense-mediated decay
OAC	oesophageal adenocarcinoma
OVs	oncolytic viruses
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PD1	programmed cell death protein 1
PDGFRA	platelet-derived growth factor receptor-α
PH	pleckstrin homology
PLA	Proximity ligation assay
RMS	Rhabdomyosarcoma
SBP	Streptavidin-Binding Peptide
SCCOT	squamous cell carcinoma of the oral tongue
SNVs	single nucleotide variants
STS	soft tissue sarcomas
SWATH-MS	Sequential window acquisition of all theoretical mass spectra
TCGA	The Cancer Genome Atlas

TMA	tissue microarray
TMT	Tandem mass tag
TSGs	tumour suppressor genes
TTSP	type II transmembrane serine-proteases
UPS	undifferentiated pleomorphic sarcoma
WES	whole exome sequence
WGS	whole genome sequence
Wt	wild type

Figures and Tables

Figures

Tables

CHAPTER ONE
Introduction

1.1 Cancer

1.1.1 Cancer background

Cancer is the second leading cause of death in the world, after cardiovascular diseases. It is not a new disease and has afflicted people throughout the world. Some of the earliest evidence of human bone cancer was found in mummies in ancient Egypt, and the disease is mentioned in ancient manuscripts dated about 1600 B.C. [1]. Cancer is a complex and heterogeneous disease to which all organs and tissues in our body are susceptible. It develops when normal cells in a particular part of the body begin to grow out of control [1, 2]. Cancer cells continue to grow, divide and re-divide instead of dying, and thus form colonies of new, abnormal, cells. Some types of cancer cells often travel to other parts of the body through blood circulation or lymph vessels, and this process is known as metastasis. For example, when a breast cancer cell spreads to the liver through blood circulation, the cancer is still called breast cancer, rather than liver cancer [1].

1.1.2 Cancer incidence

The occurrence of cancer is increasing in the world because of the growth and aging of the population, as well as an increasing prevalence of established risk factors such as smoking, alcohol consumption, and being overweight and physically inactive [3]. In 2012, the worldwide burden of cancer involved 14 million new cases per year, a figure expected to rise to 24 million annually within the next two decades [4]. Lung and breast cancers are among the most frequently diagnosed cancers and are leading causes of cancer death in men and women, respectively, both overall and in the less developed world. In more developed countries, however, prostate cancer is the most frequently diagnosed cancer among men and lung cancer is the leading cause of cancer death among women. Other frequently diagnosed cancers worldwide include those of the liver, stomach and colorectal among males, and of the stomach, cervix, uterus, and colorectal among females. In more developed countries, bladder cancer among males and uterine cancer among females are also frequently diagnosed [3].

Cancer type frequencies are different in different regions of the world. Most cancers that are frequently observed in one population are relatively rare in another, and these patterns vary over time. For example, oesophageal cancer is common among men in East Africa but rare in West Africa. Colorectal cancer, once rare in Japan, increased fourfold in incidence in just two decades [3].

1.1.3 Cancer genetics

There is accumulating evidence indicating that genetic variation accounts for a proportion of susceptibility to common diseases such as diabetes, cardiovascular disease and cancer [5]. Genetic variation in the human genome is present in many forms and occurs at different frequencies throughout the genome. The different forms of genetic variation include base substitutions, insertions and deletions (indels) of bases, rearrangements caused by breakage and abnormal re-joining of DNA, and changes in the copy number of DNA segments [5, 6].

A century of laboratory and clinical research has made it clear that cancer is fundamentally a genetic disease, driven by heritable changes in the cancer genome or chromatin, arising either in the germline or during the many subsequent steps required for a normal cell to become malignant [7]. Most common cancers are caused by acquired mutations in somatic cells. In contrast, specific germline mutations account for rare hereditary cancer syndromes [8].

1.2 A brief history of sequencing approaches in cancer

Over the past half-century, many technologies have been deployed to characterise systematically, at ever-increasing levels of resolution, the state of cancer genomes across the range of cancer types (fig 1.1) [6]. Chromosomal abnormalities were first suggested to be causally related to tumorigenesis in 1914. This notion was confirmed half a century later with the arrival of sophisticated cytogenetic techniques, which facilitated the identification of the first specific translocation driving human neoplasia: the t(9;22)(q24;q11) translocation responsible for generating the Philadelphia chromosome, resulting in the BCR–ABL fusion oncoprotein causing leukemiagenesis [9].

By the middle of the 20th century, scientists had solved the complex problems of the chemistry and biology behind cancer. Watson and Crick received the Nobel Prize in 1962 for the discovery

of the helical structure of DNA. Later, scientists learned how genes worked and how they could be damaged by mutations. Scientists found that cancer could be caused by chemicals (carcinogens), radiation, viruses, and also could be inherited from ancestors. Most carcinogens cause damage to the DNA which leads to abnormal growth of cells [1].

Cytogenetic techniques, such as fluorescence in situ hybridization (FISH) and comparative genome hybridization (CGH), helped identify several additional recurrent chromosomal aberrations, and revealed important aspects of cancer biology. In addition, early sequencing technologies developed during the same period by Ray Wu and Frederick Sanger allowed for the targeted sequencing of specific regions, and therefore helped identify recurrent mutations in genes of interest [10]. Technically, standard Sanger sequencing identifies linear sequences of nucleotides by electrophoretic separation of randomly terminated extension products [11].

In 1990, just five years after Sanger's chain termination sequencing method was partially automated for the first time, the Human Genome Project was launched as collaboration between genome centres in several countries. A draft sequence was published eleven years later, and the project was deemed complete in 2004 [10].

The Human Genome Project and its accompanying need for large-scale sequencing approaches and data analysis inspired the creation of next generation sequencing (NGS) methods, which allowed for DNA fragments to be sequenced in a massively parallel fashion [10].

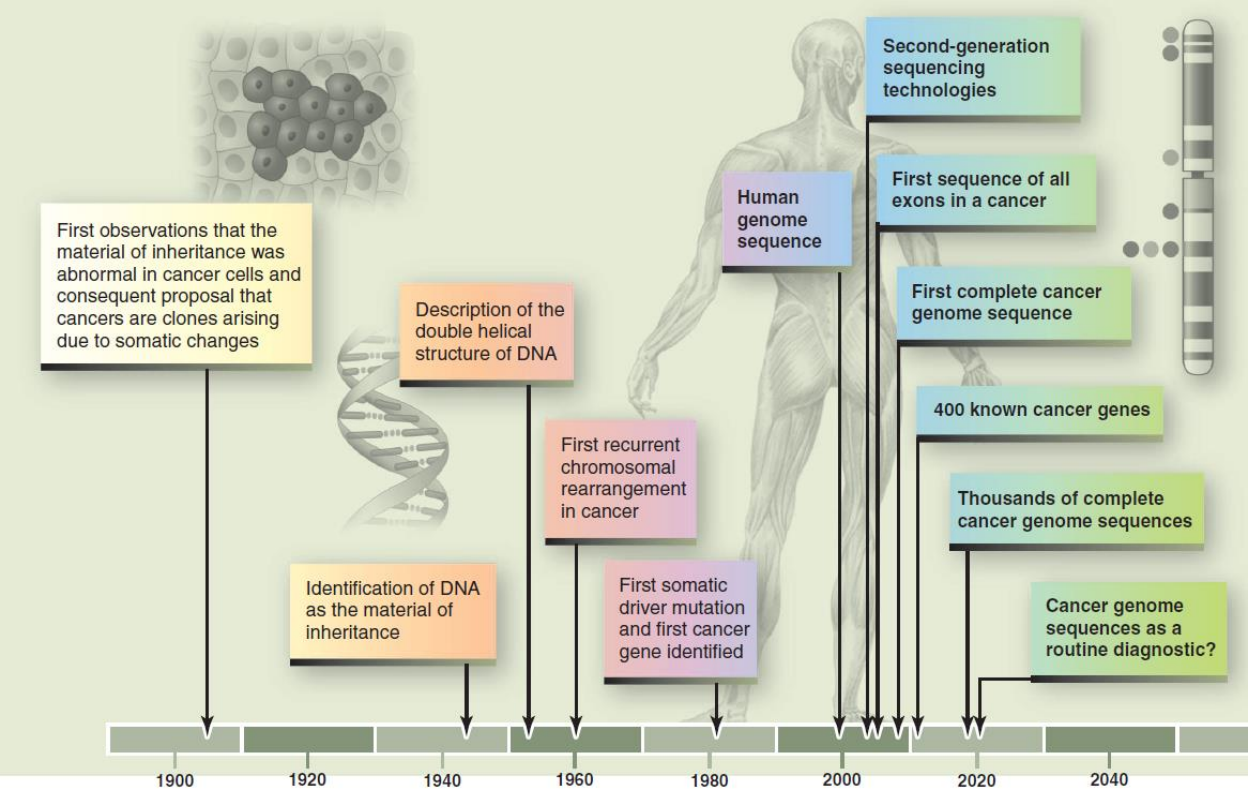


Figure 1.1. Time line showing key events in the investigation of the cancer genome [6].

1.3 Next generation sequencing (NGS) in cancer

1.3.1 NGS background

NGS refers to the high-throughput DNA sequencing technologies which are capable of sequencing large numbers of different DNA sequences in one reaction [11]. The advent of NGS technologies has stimulated rapid cataloguing of all alterations in cancer genomes, and has enabled researchers to look at large-scale genomic events such as chromosomal lesions and copy-number variations, as well as small-scale aberrations represented by point mutations, and small indels. This has provided enormous insight into the genomic landscape of several tumour types, identifying new drugable targets, and revealed the heterogeneity of many tumours [9, 12]. In 2008, the full human genome sequence was obtained using NGS technology. In the same year, the first cancer genome and the first tumour–normal genome comparison was published in a study that used data from acute myeloid leukaemia (AML). That study identified ten genes with somatic mutations, only two of which have been previously described in AML [10]. The study established that genome sequencing of tumours allowed the identification of cancer-associated

mutations, heralding a revolution in cancer research in which similar analyses across an expanding number of tumours have been published.

1.3.2 NGS workflow

NGS reads are produced from fragment 'libraries' that have not been subject to the conventional vector-based cloning and *Escherichia coli (E. coli)*-based amplification stages used in Sanger sequencing. This way, some of the cloning bias issues can be avoided [13].

The workflow to produce NGS libraries is straightforward; DNA fragments are prepared for sequencing by ligating specific adapter oligonucleotides to both ends of each DNA fragment (fig 1.2, A). The adapter sequences are capable of hybridising to the oligonucleotides on the Sequencer's flow cell surface. Cluster generation proceeds when denatured DNA libraries are allowed to randomly hybridize to the oligonucleotide lawn in the channels by their adapter ends (fig 1.2, B). A covalently-attached DNA fragment is created by extension of the flow cell oligonucleotides using the hybridized library fragment as a template. The original library strands are then denatured and washed away, leaving only the newly-synthesized strand. A complimentary copy of the covalently-bound strand is then generated through bridge amplification, a process by which the strand bends to hybridise to an adjacent and complementary oligonucleotide, thereby allowing the polymerase to extend the complementary strand. Bridge amplification is repeated 24 times to produce clusters of DNA clones in which half of the molecules represent the forward orientation and the other half the reverse orientation. Bases are read using a cyclic reversible termination strategy, which sequences the template strand one nucleotide at a time through progressive rounds of base incorporation, washing, imaging and cleavage. In this strategy, fluorescently-labelled 3´-O-azidomethyl-dNTPs are used to pause the polymerisation reaction, enabling removal of unincorporated bases and fluorescent imaging to determine the added nucleotide (fig 1.2, C). Depending on the instrument and library construction protocol used, forward and reverse reads can be paired to map both ends of linear DNA fragments during sequencing (paired-end sequencing). This process is repeated until the predetermined sequence (read) length is reached. The sequence reads obtained are then aligned

to a reference genome, using bioinformatics software, to analyse the differences between them (fig 1.2, D) [14, 15].

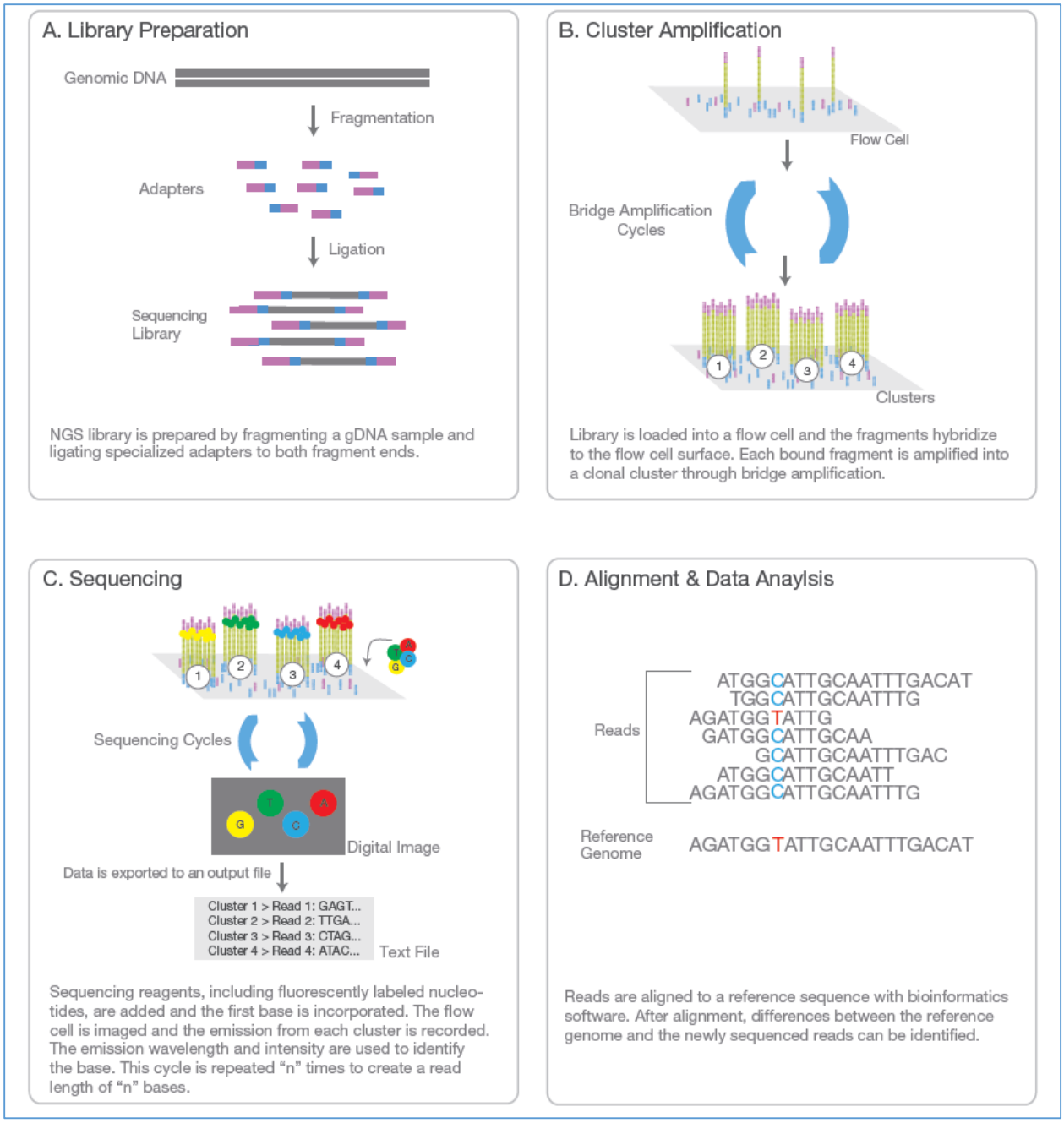


Figure 1.2. Next Generation Sequencing chemistry overview.
(http://www.illumina.com/content/dam/illuminamarketing/documents/products/illumina_sequencing_introduction.pdf)

1.3.3 Applications of NGS

NGS techniques can be broadly classified into applications for investigating the genome or transcriptome. Genomic assays include whole genome sequence (WGS), whole exome sequence (WES), and targeted re-sequencing of specific regions to discover variants associated with cell function or disease [15].

The application of NGS is becoming important for both rare disorders, and for phenotypically and genetically heterogeneous common diseases such as epilepsy and intellectual disability. For example, it was reported that a patient with epilepsy, intellectual disability, and other features, had undergone more than 20 different genetic tests before the disease-causing mutation was found in the *SCN2A* gene by WES [16].

NGS has been applied to great effect in cancer. The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) have performed WGS and WES on thousands of tumour–normal pairs. These studies have described the mutational landscapes for over 20 cancer types, showing that tumours can vary in both the type and number of mutations. These global descriptions have been integral to the development of background mutation rates that are necessary for the detection of cancer driver genes [14].

1.3.4 WGS and WES

WGS covers the whole genome and provides a unique window by which to investigate genetic or somatic variations, leading to new avenues for exploration of normal and disease phenotypes. However, the massive quantity of data and the requirement of significant computational resources make WGS cost prohibitive for routine genetic and biological studies. In contrast, WES focuses on capturing and sequencing protein-coding regions (exomes), which cover between 1 and 2% of the genome, limiting the data to a more functionally informative part of the genome, and thus has become a popular choice for genetic studies to identify coding and splice-site variants from a large number of samples in a short time period [15].

WES is currently the popular technology for sequencing cancer genomes, and has led to an abundance of discoveries in many cancer types [17].

WES is a relatively comprehensive, inexpensive, and rapid way of identifying coding mutations, compared with other variant detection methods. However, there are still some limitations. None of the WES capture probe sets seem to target all of the exons, so some exons would be missed. The capture step is not uniform and tends to have a bias against high GC-rich regions. The capture would also skip as yet unidentified exons, and so miss variants [15].

1.3.5 Transcriptome sequencing

The development of cancer is a multistep process that involves the accumulation of a wide range of genetic and phenotypic alterations, leading to the aberrant expression of genes that regulate cell proliferation. Microarrays have been used to identify differential gene expression profiles in many cancer types, which have aided the screening, prognosis, and classification of tumours. However, microarrays have limitations, including low sensitivity and hybridization artefacts [18].

The development of NGS has also revolutionized transcriptomics by enabling RNA analysis through the sequencing of complementary DNA (cDNA). This method, termed RNA sequencing (RNAseq), has eliminated many challenges posed by hybridization-based microarrays and Sanger sequencing-based approaches that were previously used for measuring gene expression, and it has revolutionized our understanding of the complex and dynamic nature of the transcriptome [19]. RNAseq is a mode of deep sequencing that enables evaluation of a complete set of an organism's transcribed genes, or transcriptome, and noncoding RNAs such as micro-RNAs [20].

A typical RNAseq experiment consists of isolating RNA, converting it to cDNA, preparing the sequence library, and sequencing it on an NGS platform [19].

RNAseq is performed to measure gene expression, detect fusion transcripts, and for splicing analysis. It can also provide an observation of the underlying tumour DNA sequence, via transcription. Therefore, it can confirm the expression of the mutations detected in the genome [17].

There are many advantages of RNAseq: rapid, precise and quantitative measurement of gene expression; high sensitivity allows detection of low-abundance transcripts; enables identification of transcripts; facilitating the discovery of single nucleotide polymorphisms and rare mutations; and the detection of previously unrecognized genes [20].

Despite its advantages in cancer studies, RNAseq has some limitations. Accurate sequence annotation and data interpretation can be computationally challenging, transcript quantitation can be affected by biases introduced during cDNA library construction and sequence alignment, and its cost remains prohibitive for many laboratories [20].

1.4 Cancer driver genes

1.4.1 Oncogenes and tumour suppressor genes (TSGs)

The advanced genomic studies of different cancer types have enhanced the accuracy and coverage of the identification of cancer-related genes that could drive or protect cancer development. The majority of gene variants found in cancer fall into two categories: gain-of-function mutations in proto-oncogenes, which stimulate cell growth, division, and survival; and loss-of-function mutations in tumour suppressor genes (TSGs) that normally help prevent unrestrained cellular growth and promote DNA repair and cell cycle checkpoint activation [21]. Oncogene mutations are usually dominant; such that one mutant allele is enough to start switching on a cellular activity. TSG mutations tend to be recessive, so that two copies of the gene must be mutated to cause loss of function. However, recent studies show that even partial inactivation of TSGs could critically contribute to tumorigenesis [22]. The oncogenes and their related pathways are promising avenues for developing novel drugs, including antibodies and small synthetic molecules.

Abnormal activation of proto-oncogenes can occur by chromosomal translocation resulting in fusion genes such as BCR–ABL in chronic myelogenous leukaemia (CML), by point mutations resulting in an altered protein product of the gene, and by amplification resulting in many copies of the oncogene, for example the amplification of the *HER2* gene, which codes for a growth factor receptor, in many cases of breast cancer [2].

TP53 is one of the most interesting TSGs, which codes for a 53 KDa protein. This protein interacts with many cellular genes involved in cellular growth control, and its main function is to halt the cell division cycle in cells with DNA damage, to induce either DNA repair or programmed cell death. Cells in which this function of ‘cellular keeper’ for growth and division TP53 is

compromised can be more readily accumulate mutations and may more rapidly progress to overt malignancy [2].

1.4.2 Somatic mutations

Somatic mutations are the variants found in the genomes of tumour cells, but not in their matched normal cells. They often play important roles in tumour initiation, progression and metastasis. They can be single nucleotide variants (SNVs) that change the amino acid sequence (missense mutation), or prematurely truncating encoded proteins (nonsense mutation), or indels that delete or insert some amino acids and disrupt the protein function. Discovering such cancer-related variants is confounded by the presence of millions of germline mutations. To distinguish somatic from germline variants, it has become important to sequence matched tumour–normal samples from the same patient [23].

Somatic mutations are thought to occur in the genomes of all normal cells as they proceed through the rounds of cell division that take place during development in utero, and during replenishment of body tissues in postnatal life. Additional somatic mutations continue to accumulate in cancer cells as they divide. The rate and types of these mutations can be increased by exogenous and endogenous exposures that cause DNA damage and are mitigated by DNA repair processes [6].

1.4.2.1 Driver and passenger mutations

Although most of the somatic mutations that steadily accumulate in our cells are harmless, occasionally a mutation affects a gene or regulatory element and leads to a phenotypic consequence. A fraction of these mutations can confer a selective advantage to the cell, leading to preferential growth or survival of a clone. These types of mutations are called driver mutations. The other types of mutation have no phenotypic or biological effect and are called passenger mutations [24].

The number of driver mutations in a cancer cell reflects the number of mutated cancer genes and thus the deregulation of cell biological processes required to convert a normal cell into a symptomatic cancer clone [6].

1.4.2.2 Mutation timing

Tumours evolve from benign to malignant lesions by acquiring a series of mutations over time. The first mutation provides a selective growth advantage to a normal cell, allowing it to outgrow the cells that surround it and become a microscopic clone. A study of colorectal cancer has shown that the first mutations occur in the *APC* gene (fig 1.3). The small adenoma that results from this mutation grows slowly, but another mutation in another gene, such as *KRAS*, unleashes a second round of clonal growth that allows an expansion of cell number. This process of mutation followed by clonal expansion continues, with mutations in genes such as *PIK3CA* and *TP53*, eventually generating a malignant tumour that can invade through the underlying basement membrane and metastasise to lymph nodes and distant organs [25]. Thus it is important to consider the cancer stage when analysing the cancer genome to detect somatic mutations.

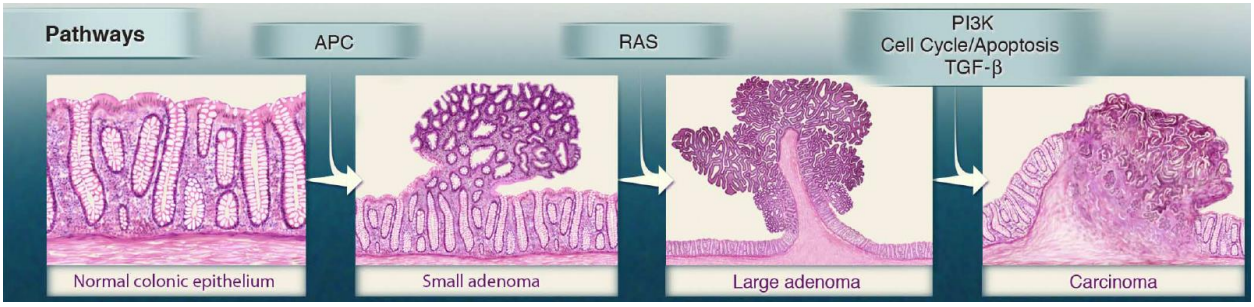


Figure 1.3. Genetic alterations and the progression of colorectal cancer*. The major signalling pathways that drive tumorigenesis are shown at the transitions between each tumour stage. One of several driver genes that encode components of these pathways can be altered in any individual tumour [25].*

1.4.2.3 Mutation signature

All cancers develop as a consequence of accumulating somatic mutations in their genomes, which may be a result of the intrinsic slight infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA, or defects in DNA repair. Different mutational processes often generate different combinations of mutation types, termed mutation signatures [26]. Thus, understanding the mutation signature of each tumour type can provide insights into the forces that cause somatic mutations. The impact of NGS in understanding cancer mutation signatures became clear in 2010 when two studies reported the pattern of somatic mutations in a malignant melanoma and small cell lung carcinoma. They found that there was a strong signature of tobacco carcinogens in the genome of the lung cancer, while the mutational signature of ultraviolet radiation overwhelmed the melanoma genome [27, 28].

In lung cancer, it was reported that tobacco smokers have an average 10-fold increase in the burden of somatic mutations in their cancer genomes compared to non-smokers. Consistent with the experimental evidence for tobacco carcinogens, this elevation is mainly due to the increase of the mutant of C>A transversions [27].

Alexandrov et al. 2013, [26] analysed 4,938,362 mutations, substitutions and indels, from 7,042 primary cancers of 30 different classes. They extracted more than 20 distinct mutational signatures (fig 1.4, A), and found that some of these signatures are present in many cancer types (fig 1.4, B). They also found that certain signatures are associated with age of the patient at the time of diagnosis, such as signatures 1A and 1B, which may have been caused by an endogenous mutational process. Other signatures may have been caused by exposure to exogenous mutagenesis. Signature 4 is found in cancers associated with tobacco smoking and has the mutational features associated with tobacco carcinogens. Signature 7 is found in malignant melanoma and squamous carcinoma of the head and neck and has the known features of UV light-induced mutations.

The ability to detect the mutation signature of a specific tumour would be expected to have an impact on cancer prevention by conclusively identifying risk factors. Furthermore, ongoing mutational processes in tumours may reflect actionable features, influencing prognosis and treatment.


A
C>A C>G C>T T>A T>C T>G
Signature 1A
Signature 1B
Signature 2
Signature 3
Signature 4
Signature 5
Signature 6
Signature 7
Signature 8
Signature 9
Signature 10
Signature 11
Signature 12
Signature 13
Signature 14
Signature 15
Signature 16
Signature 17
Signature 18
Signature 19
Signature 20
Signature 21
Percentage of mutations
D Dinucleotide mutations
T Transcriptional strand bias
I Indels
B
ALL
AML
Bladder
Breast
Cervix
CLL
Colorectum
Oesophagus
Glioblastoma
Glioma low grade
Head and neck
Kidney chromophobe
Kidney clear cell
Kidney papillary
Lung adenocarcinoma
Liver
Lung small cell
Lung squamous
Lymphoma B cell
Medulloblastoma
Melanoma
Myeloma
Neuroblastoma
Pilocytic astrocytoma
Pancreas
Ovary
Prostate
Stomach
Thyroid
Uterus
Prevalence in cancer samples
Probable association
Validated mutational signatures
Signature 1A 11.7% Age
Signature 1B 60.7% Age
Signature 2 14.4% APOBEC
Signature 3 9.9% BRCA1/2 mutations
Signature 4 12.1% Smoking
Signature 5 14.4%
Signature 6 2.6% DNA MMR deficiency
Signature 7 5.0% Ultraviolet light
Signature 8 2.0%
Signature 9 0.6% Immunoglobulin gene hypermutation
Signature 10 0.5% Pol ε mutations
Signature 11 0.6% Temozolomide
Signature 12 1.4%
Signature 13 2.2% APOBEC
Signature 14 0.1%
Signature 15 0.5%
Signature 16 1.1%
Signature 17 1.8%
Signature 18 2.2%
Signature 19 0.2%
Signature 20 0.5%
Signature 21 0.3%
Other 13.6%
Mutational signature present
Total validated mutational signatures in a cancer type
Total cancer types in which a signature is operative

Figure 1.4. Validated mutational signatures found in human cancer and the presence of mutational signatures across human cancer types*. **A,** each signature is displayed according to the 96 substitution classification defined by the substitution class and sequence context immediately 39 and 59 to the mutated base. The probability bars for the six types of substitutions are displayed in different colours. The mutation types are on the horizontal axes, whereas vertical axes depict the percentage of mutations attributed to a specific mutation type. All mutational signatures are displayed on the basis of the trinucleotide frequency of the human genome. Asterisk indicates mutation type exceeding 20%. **B**, Cancer types are ordered alphabetically as columns whereas mutational signatures are displayed as rows. Prevalence in cancer samples indicates the percentage of samples from the data set of 7,042 cancers in which the signature contributed significant number of somatic mutations. For most signatures, significant number of mutations in a sample is defined as more than 100 substitutions or more than 25% of all mutations in that sample. MMR, mismatch repair [26].*

1.5 Signalling pathways in cancer

All of the driver genes detected in different cancers can be classified into one or more of 12 pathways (fig 1.5). These pathways can themselves be further organized into three core cellular processes: cell fate, cell survival, and genome maintenance [25].

Figure 1.5. Cancer cell signalling pathways and the cellular processes they regulate*. All of the driver genes can be classified into one or more of 12 pathways (middle ring) that confer a selective growth advantage (inner circle). These pathways can themselves be further organized into three core cellular processes (outer ring) [25].*

1.6 Cancer immunotherapy

1.6.1 Neoantigens and the immune system

Human cancers arise through genetic changes that facilitate cellular immortality, but at the same time create foreign antigens, the so-called neoantigens, which should render neoplastic cells detectable by the immune system and target them for destruction. Nevertheless, although the immune system is capable of noticing differences in protein structure at the atomic level, cancer cells manage to escape immune recognition and subsequent destruction [29].

Immunotherapies that boost the ability of endogenous T cells to destroy cancer cells have demonstrated therapeutic efficacy in a variety of human cancers. Until recently, evidence that the endogenous T cell compartment could help control tumour growth was, in large part, restricted to preclinical mouse tumour models and to human melanoma [30].

The field of cancer immunotherapy has recently received a significant boost, encouraged primarily by the approval, in 2010, of the autologous cellular immunotherapy, sipuleucel-T, for the treatment of prostate cancer, and the approval of the anti-cytotoxic T lymphocyte-associated protein 4 (CTLA-4) antibody, ipilimumab, and also of anti-programmed cell death protein 1 (PD1) antibodies for the treatment of melanoma in 2011 and 2014, respectively [29].

The clinical success of immune checkpoint antagonists targeting the PD-1 pathway in multiple tumour types provides clear evidence that harnessing the immune system can lead to durable tumour regressions in cancer patients [31].

1.6.2 Tumour specific antigens

The recent development of NGS, along with advances in bioinformatics, have enabled systemic analyses of the mutation load of the tumour as well as identification of the potentially immunogenic neoantigens [32].

T cells are able to reject tumours on recognition of tumour-specific antigens bound to the major histocompatibility complex (MHC) molecules of tumour cells. MHC is a protein complex that presents antigens on the cell surface. In humans, the MHC is encoded by the human leukocyte antigen (HLA) gene locus. Antigens with high tumour specificity, that is displayed by tumour cells but not by normal cells, have the potential to elicit a tumour-specific immune response,

minimizing the risk of adverse side effects, and are therefore of great interest for cancer immunotherapies such as engineered T cells and therapeutic vaccines [33].

Neoantigens can be considered to be strictly tumour-specific because they originate from the expression of mutated genes that are present in malignant cells but not in the normal tissues. To elicit an immune response, the mutated proteins must be proteolytically processed into short peptides and then bound to MHC molecules, to be presented to T cells (Fig 1.6, A). When NGS data are available from matched tumour and normal tissues, neoantigens can be predicted in *silico* by integrating three computational tasks (fig 1.6, B): the identification of mutated proteins from matched tumour–normal samples, followed by HLA typing, and then prediction of neoantigens–MHC binding affinity [33].

A large fraction of the mutations in human cancers is not shared between patients at meaningful frequencies and may therefore be considered patient-specific. Because of this, technologies to interrogate T cell reactivity against putative mutation-derived neoantigens need to be based on the genome of an individual tumour [34]

1.6.3 Classes of neoantigens

There are three classes of antigens that have high tumour specificity: the first is viral antigens, which are derived from viral genes expressed in virus-infected tumour cells, second is cancer germline antigens, which are proteins that are normally expressed only by trophoblasts and germline cells but that have aberrant expression in tumour cells, and third is neoantigens, which are peptides that arise from the expression of somatically mutated genes [33].

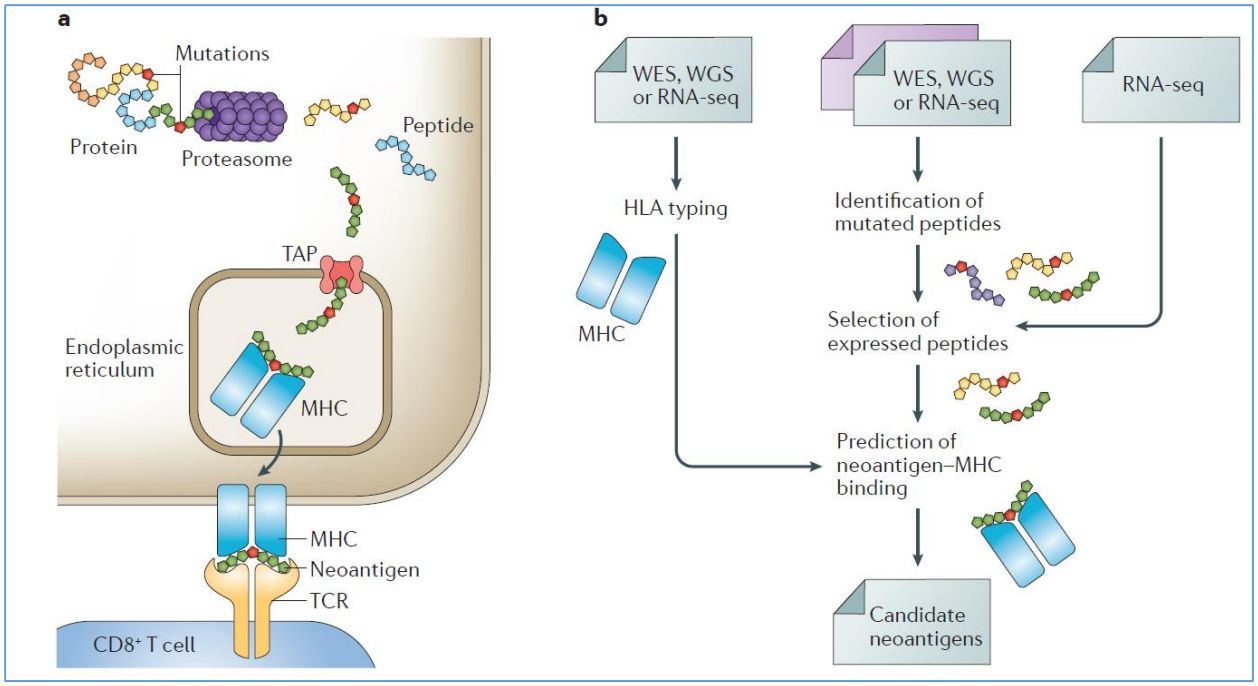


Figure 1.6. Identification of cancer neoantigens. A, *neoantigens originate from mutated proteins expressed in cancer cells. The mutated protein is cleaved into shorter peptides by the proteasome and transported by transporter associated with antigen processing (TAP) to the endoplasmic reticulum, in which the peptides bind the major histocompatibility complex (MHC) molecule. Then, the peptide–MHC complex is displayed on the cell surface of the antigen-presenting cell to be recognized by the T cell receptor (TCR) of CD8$^+$ T cells.* ***B,*** *prediction of candidate neoantigens from next-generation sequencing (NGS) data requires the implementation of several computational tasks: prediction of mutated peptides from whole-exome sequencing (WES); whole-genome sequencing (WGS) or RNA sequencing (RNAseq) data from matched tumour–normal samples; selection of expressed peptides by integrating RNAseq data of the tumour sample; human leukocyte antigen (HLA) typing from WES, WGS or RNAseq data of the tumour sample; and prediction of peptide–MHC binding for specific HLA alleles [33].*

1.6.4 Different strategies to activate immune cells

Strategies to activate effector immune cells include vaccination with tumour antigens or augmentation of antigen presentations to increase the ability of the patient's own immune system to mount an immune response against neoplastic cells [6]. Additional stimulatory strategies encompass adoptive cellular therapy (ACT) in an attempt to administer immune cells directly to patients, the administration of oncolytic viruses (OVs) for the initiation of systemic antitumor immunity, and the use of antibodies targeting members of the tumour necrosis factor receptor superfamily, so as to supply co-stimulatory signals to enhance T cell activity [29].

1.6.4.1 Vaccines targeting individual mutations

Mutation-based vaccination attempts represent 'off-the-shelf' approaches, as they target single or multiple frequently shared neoantigens, such as mutant RAS and mutant epidermal growth factor receptor (EGFR). These were used to stratify patients according to the presence or absence of the respective mutations in their tumour. Various attempts have been made to apply neoantigen-based vaccine strategies in mouse models. Vaccination studies with synthetic peptides and mRNA-encoding mutations identified by NGS in three different mouse tumour models revealed that a significant portion of non-synonymous point mutations is immunogenic. Developed mRNA vaccines that encoded single MHC class II neoepitopes were capable of controlling the growth of established mouse melanoma and colon cancer tumours [35].

Ongoing clinical trials have provided evidence for the feasibility of individualized vaccination of cancer patients with unique mutations. Once neoepitopes have been selected, the patient's individualized vaccine can be manufactured. These vaccines can be in the form of synthetic peptides and antigen encoding DNA or RNA (fig 1.7) as formats which are able to deliver effective mutant MHC class I and class II neoepitopes [35].

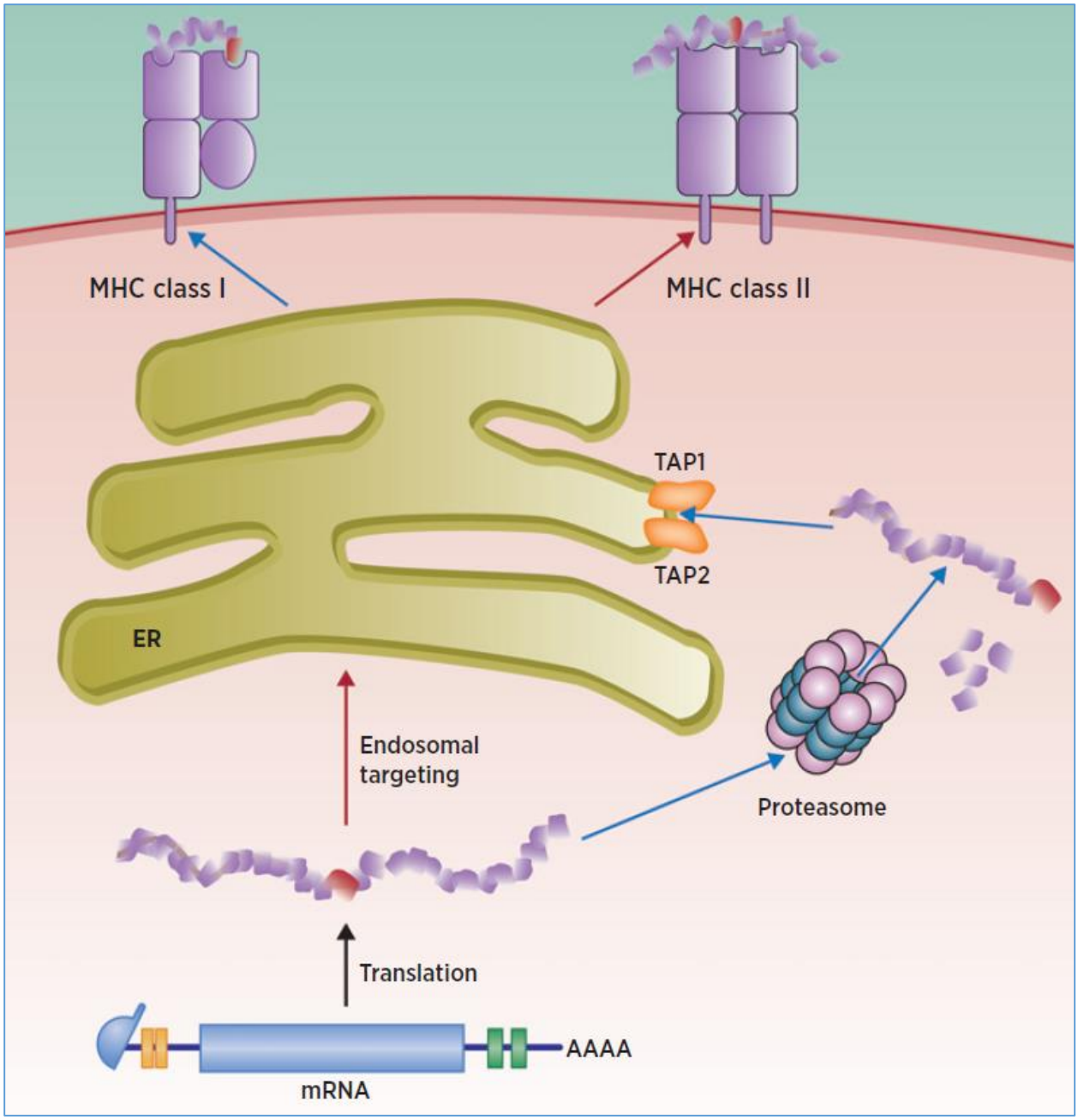


Figure 1.7. Schematic mechanism of an mRNA-based neoepitope vaccine*. The mRNA vaccine-encoding parts of the neoantigen are translated in the cytosol. Subsequently, the mutated peptides are C-terminally truncated by the proteasome and transported into the endoplasmic reticulum (ER), where loading of the neoepitope on MHC class I takes place. Routing of mutated peptides into the endosomal pathway allows the presentation by MHC class II molecules [35].*

1.7 Brief Aims of each Results chapter

1.7.1 Detecting somatic mutations in undifferentiated pleomorphic sarcoma (UPS)

Sarcomas are an intriguingly complex set of tumours, arising from mesenchymal tissue and representing approximately 1% of all adult cancers. One of the more perplexing features of soft tissue sarcomas (STS) is their heterogeneity. This fascinatingly diverse group boasts over 50 distinct subtypes, each one unique with regard to histology and clinical management. This complexity has made it difficult to advance our molecular understanding of these tumours. Patient samples are rare, and thus researchers have been forced to combine sarcomas from multiple different subtypes when conducting genomic studies. Although some sarcomas, such as gastrointestinal stromal tumours, are defined by a simple genetic event that allows for treatment with rational therapeutics, other STS with less-well defined mutations are in need of a deeper molecular understanding to identify new drug targets. Of particular note is undifferentiated pleomorphic sarcoma (UPS), the most common STS diagnosed in adults. Similar to many other STS cases, the majority of patients with UPS are treated with surgical resection and radiotherapy. In the case of metastatic disease, patients with UPS receive a combination of doxorubicin and/or ifosfamide-based therapy, which is both toxic and demonstrates only marginal efficacy. Classically, UPS is a diagnosis of exclusion, defined by a lack of specific histological or molecular features. UPS is extremely unstable at the genomic level. The few genomic analyses of UPS conducted to date have identified dysregulation of cell cycle and tumour suppressor pathways. However, to the best of our knowledge, these studies did not report any disruptions in signal transduction pathways that may provide new molecular targets for UPS [36].

1.7.2 WES of 20 UPS tumour-normal tissues

We aimed to use newly available software to define the mutated genomic landscape using UPS as the model cancer. In an attempt to find a drug target and specific biomarker in UPS, we have sequenced and analysed the WES of 20 UPS tumours and their matched normal adjacent tissues. We have detected a high number of somatic mutations in each patient using the Biomedical Cancer Research Workbench of the CLC-bio software. We have selected the highly mutated genes in these patients and studied their pathways.

1.7.3 RNAseq analysis of three UPS

RNA sequencing is emerging as being used as another landscape to identify and stratify mutated and expressed genes. However, the currently available RNAseq software tools do not share a high degree of integrity and we have used newly available software to quantify expressed mutated RNA landscapes. We have also sequenced the whole RNA (transcriptome) of three UPS tumour tissues and their matched normal tissues. As not all the somatic mutations detected in the DNA would be expressed in the RNA, we have compared the tumour RNAseq to the tumour DNAseq to detect expressed somatic mutations at the RNA level. This step is necessary in order to select the mutation for further studies, but it has limitations as not all genes would be expressed at the time of tumour resection and we would miss some genes as mentioned above. The tumour RNAseq of the three UPS patients were also compared to the RNAseq of matched normal tissues to detect genes with differential expression in tumour and normal tissues of UPS.

1.7.4 Analysing WES from small sets of patients to define cut-off parameters in CLC-bio software

Before we started the analysis of the UPS data search using the CLC-bio software, we have applied WES of tumour tissues and two matched normal tissues, normal adjacent tissue and germline blood, from a small set of patients with head and neck cancer (tongue cancer) to define cut-off parameters in the use of the new software. Patients were divided into two groups: the first group were young patients, and the second group were old patients. All patients have no history of smoking, alcohol consumption or viral infection, the most known risk factors for tongue tumours. Although the number of patients was low (three young and two old), we have analysed their WES using the CLC-bio software, and detected somatic mutations in the tumour tissues and compared them to the normal adjacent tissue/blood. This analysis has helped us to set up the parameters for the UPS data analysis.

We were also provided with the RNAseq of four patients (two young and two old), and these were used in the analysis to get an idea of the percentage of the expressed somatic mutations detected in the genome.

1.8 Genetics of oesophageal adenocarcinoma (OAC)

During the course of my studies, OAC has emerged as a progression model for genomics since intermediates in the disease can be collected. In this model, there are novel driver mutations identified of unknown function and I developed methods that can identify functions to such orphan, mutated proteins. Although, by the end of my thesis, I had no time to apply the method to UPS target proteins, the principle methodology is established by my thesis to applied to other cancer proteins. There are dozens of different cancer types. I will focus below on a cancer of unmet clinical need where genomics has shed light on mechanisms of development. OAC, including cancers of the gastro-oesophageal junction, represent a substantial health concern in Western countries due to its increasing incidence and poor prognosis. Recent genome-wide sequencing projects have shown that OAC is one of the most highly-mutated solid cancers, with a high degree of heterogeneity [37].

1.8.1 Epidemiology and risk factors of OAC

In the early to mid-twentieth century, primary OAC was rare. However, the incidence of OAC has increased 600% over the last 30 years in Western countries. The increase in incidence of OAC during this time was greater than that of any other major cancer. OAC mainly affects older white men; it is five times more common in Caucasian than in African Americans, eight times more common in men than in women, and peaks in incidence during the seventh and eighth decade of life [38, 39].

The reasons behind this rapid increase in OAC incidence are not completely understood. It is hypothesized that the rise in the prevalence of both obesity and gastroesophageal reflux disease (GERD), and perhaps the simultaneous decline in the prevalence of *Helicobacter pylori* (*H. pylori*) infection, may help explain the rise in incidence of OAC [38].

Previous studies have shown that the risk of OAC was almost eight times higher in patients with symptomatic GERD than in controls. Acid exposure in patients with GERD can damage the squamous epithelial lining of the distal oesophagus, and this damage can be followed by an aberrant healing process that results in the replacement of normal squamous mucosa by specialized, metaplastic columnar intestinal epithelium, a condition commonly known as

Barrett's oesophagus. It is thought that most OAC develop from an area of Barrett's metaplasia [38]. However, only few patients with Barrett's oesophagus develop OAC.

Infection with *H. pylori* appears to have an inverse association with OAC, with a reported 41% decrease in the risk of OAC in individuals infected with *H. pylori*. Whether this relationship is truly causal, however, remains a matter of debate. It is thought that *H. pylori* protects against Barrett's oesophagus and OAC by leading to atrophic gastritis and reduced gastric acid secretion, which in turn decreases acid exposure in the oesophagus [38].

Tobacco use has been reported to increase the risk of OAC by about twofold, and this risk is directly associated with the amount and duration of cigarette smoking [38].

1.8.2 Genetic variations of OAC

Recent efforts have been directed towards understanding the genetic landscape of OAC with a view to the identification of potentially targetable driver mutations. These studies have revealed the complexity of the genomic aberrations in OAC. The progression from Barrett's oesophagus to invasive cancer is characterized by early chromosomal instability, maybe due to p53 loss, often including genome doubling and high frequency of chromothripsis events resulting in considerable genetic diversity, followed by a later acquisition of driver mutations at sub-clonal frequencies [40].

Dulak et al [41] was the first genome-wide association study to investigate gene variations together with pathway abnormalities in OAC. In this study, 149 OAC tumour–normal pairs were subjected to WES, 15 of which had also been subjected to WGS. The study detected the highest mutation signature A>C in OAC in non-coding areas, and within coding areas they were overrepresented in less expressed genes. The study also identified 26 significantly mutated genes, the most significant of which were two known OAC tumour suppressors, *TP53* and *CDKN2A*. Many signal transduction pathways were affected by these mutations (fig 1.8). The novel significantly mutated genes included chromatin-modifying factors and regulators of invasion and motility: *SPG20, TLR4, ELMO1*, and *DOCK2*. Functional analysis of *ELMO1* detected mutations revealed an increased cellular invasion. Given that OAC is a highly invasive tumour prone to early metastasis, alterations in the RAC1 pathway may be a contributor to OAC tumorigenesis [39].

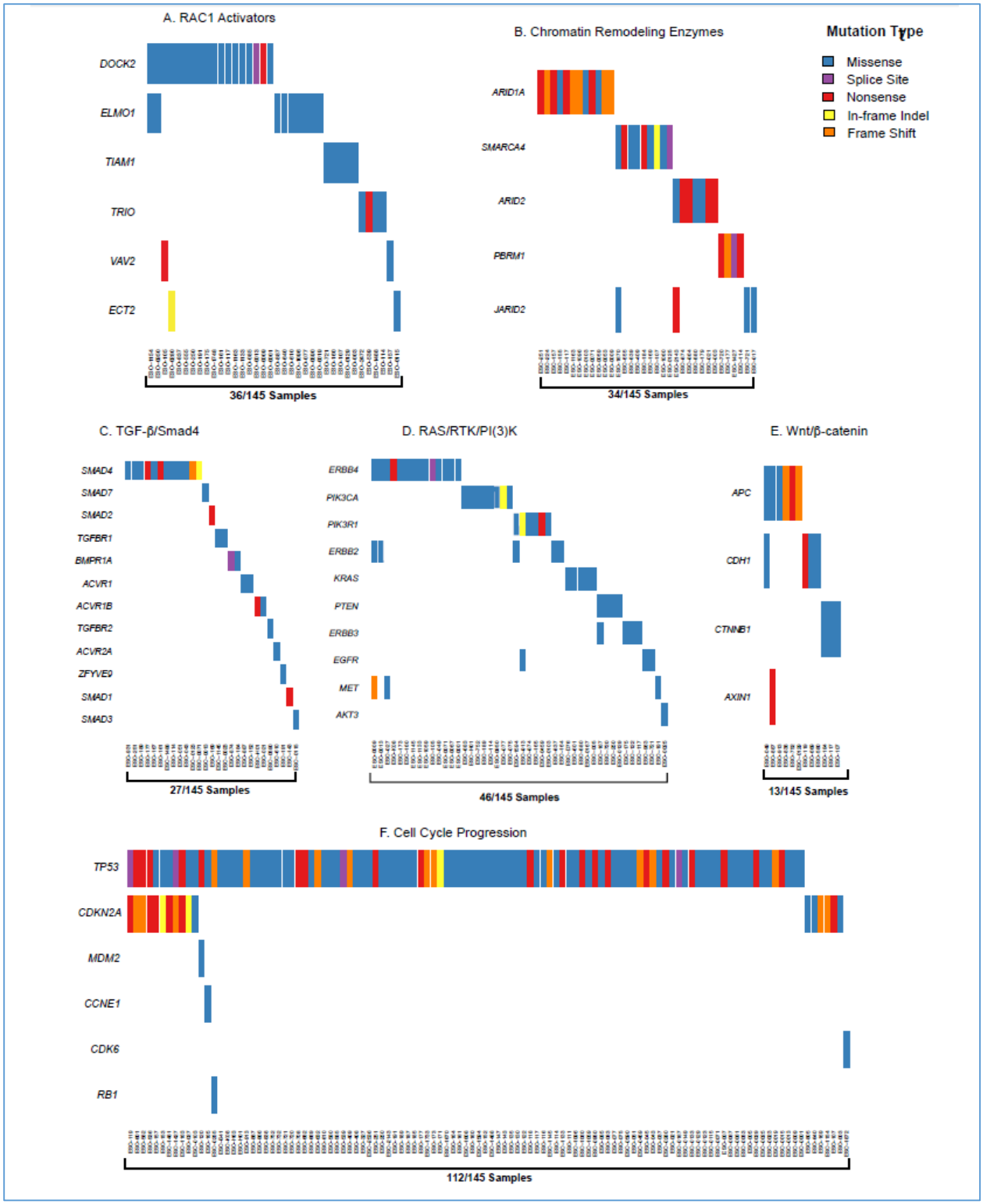


Figure 1.8. Overlap of protein coding modifications in known cancer signal transduction pathways*. **A,** RAC1 activators. **B,** Chromatin-remodelling enzymes. **C,** TGFB1(TGF-ß)/SMAD4. **D,** KRAS/PI3K/RTK. **E,** WNT/CTNNB1 (ß-catenin). **F,** Cell cycle progression [41].*

1.8.2.1 Mutation spectrum in Barrett's oesophagus and OAC

The identification of causative mutations occurring early in the OAC pathogenesis is important in order to develop clinically useful biomarkers. The mutations occurring at disease stage boundaries would be the most informative. A previous study has performed WES on 11 OAC samples and 2 samples of Barrett's oesophagus adjacent to the cancer. They reported that the majority of mutations were detected in apparently normal Barrett's oesophagus [39].

Weaver et al. [42] has analysed WGS and amplicon resequencing of 112 OAC cases, and screened their panel for the most recurrently mutated genes in samples from different stages of carcinogenesis: 90 OAC; 66 never-dysplastic Barrett's oesophagus; and 43 high-grade dysplasia (HGD) cases. They found that the most prevalent mutations in OAC were also present at a similar frequency in HGD and never-dysplastic Barrett's oesophagus samples, including mutations within cancer-associated genes, such as: *ARID1A* and *SMARCA4.* Only mutations in *TP53* and *SMAD4* were confined to HGD and OAC cases, making them good markers for malignant progression risk (fig 1.9) [39, 42]. The frequency of *SMAD4* mutation is low in OAC cases (13%) and these mutations were specific to OAC but not HGD samples, so the study has focused on *TP53* in clinical applications to differentiate between patients with HGD and those without dysplasia [39].

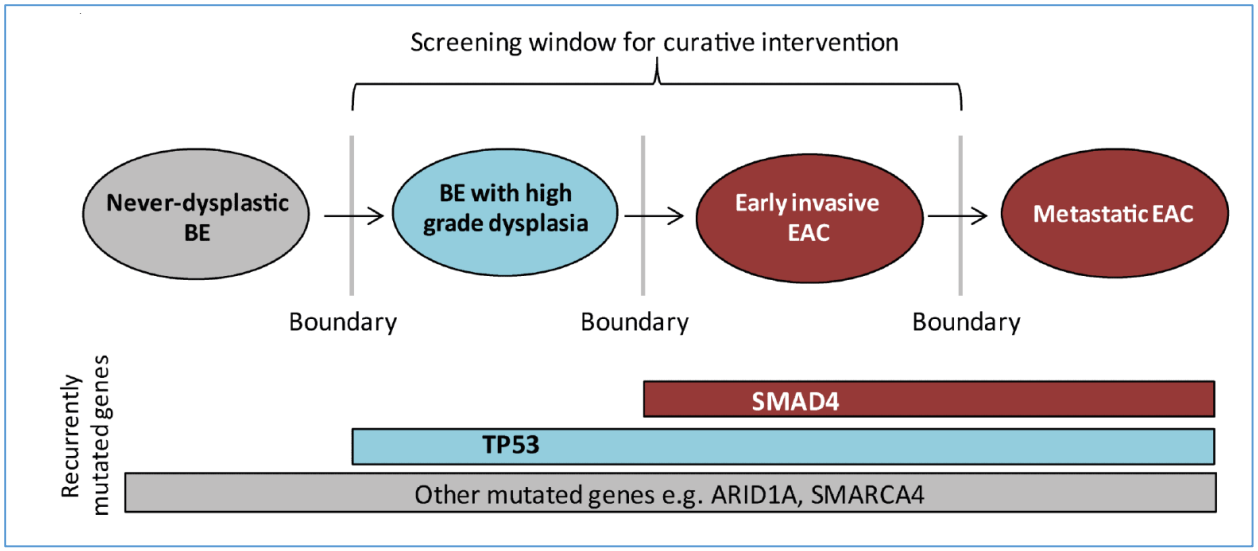


Figure 1.9. TP53 and SMAD4 mutations accurately define the boundaries in the progression towards cancer whilst other mutations appear to occur independent of disease stage. *A proposed model for the boundary-defining mutations in Barrett's oesophagus carcinogenesis. The hashed box depicts multiple other mutations that may occur and provide selective advantage at any stage of disease [42].*

1.8.3 Genome sequencing and mutation spectrum in OAC cell lines

The number of OAC cell lines available for research is limited and their genome has been only partially characterized. The availability of an accurate annotation of their mutational landscape is necessary for accurate experimental design and correct interpretation of genotype-phenotype findings. Contino et al. [37] has performed high coverage, paired and WGS on nine OAC cell lines:ESO26; ESO51; FLO-1; JH-EsoAd1; OACM5.1 C; OACP4 C; OE33; SK-GT-4 and CP-D. They have identified SNVs and indels in these cell lines by comparison with the human reference genome and known single nucleotide polymorphisms.

In order to investigate how closely these cell lines reflect the spectrum of mutations observed in human specimens, they analysed the mutational landscape of known cancer and putative OAC driver genes and compared them to the OAC mutations previously reported by Dulak et al.[41], and Weaver et al. [42]. These studies showed that 69% of OAC cases have *TP53* mutations, while all the cell lines carried at least one deleterious *TP53* mutation (fig 1.10 A, B). Two cell lines out of nine, ESO26 and JH-EsoAd1, have mutations in *SMAD4* gene, which is consistent with 13% observed in OAC (fig 1.10 A, B). *ELMO1* was also mutated in two cell lines: ESO51 and SK-GT-4.

There is no mutation detected in some genes such as *ARID1A* and *DOCK2*, which were reportedly mutated in OAC cases.

In another study, Garcia et al. [40] performed WGS on OAC tumour and leucocytes, and RNAseq on three OAC cell lines: MFD-1, FLO-1 and OE33. They have reported that the parental tumour and MFD-1 cell line carried four somatically acquired mutations in three recurrent mutated genes in OAC: *TP53, ABCB1* and *SEMA5A*, not present in FLO-1 and OE33. FLO-1 and OE33 had no expression of *ABCB1* and *SEMA5A* genes. The detection of *TP53* mutation in MFD-1 suggests that this cell line is representative of primary OAC tumours. The study has suggested that the MFD-1 cell line may be the best model of OAC to study the effects of OAC mutations.

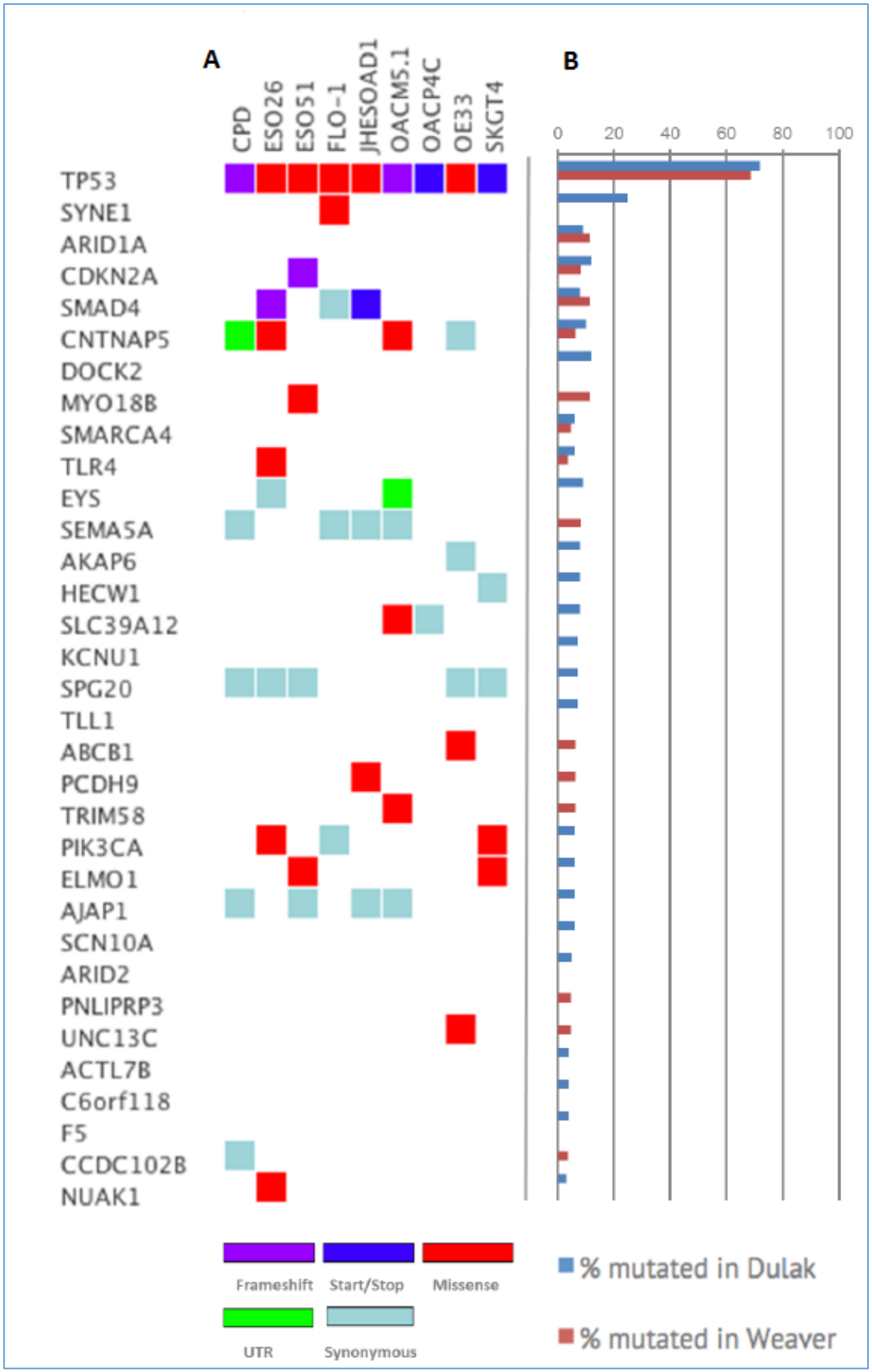

A
B
CPD
ESO26
ESO51
FLO-1
JHESOAD1
OACM5.1
OACP4C
OE33
SKGT4
TP53
SYNE1
ARID1A
CDKN2A
SMAD4
CNTNAP5
DOCK2
MYO18B
SMARCA4
TLR4
EYS
SEMA5A
AKAP6
HECW1
SLC39A12
KCNU1
SPG20
TLL1
ABCB1
PCDH9
TRIM58
PIK3CA
ELMO1
AJAP1
SCN10A
ARID2
PNLIPRP3
UNC13C
ACTL7B
C6orf118
F5
CCDC102B
NUAK1
0
20
40
60
80
100
Frameshift
Start/Stop
Missense
UTR
Synonymous
% mutated in Dulak
% mutated in Weaver

Figure 1.10. Analysis SNV of putative OAC genes identified in Dulak et al. (2013) and Weaver et al. (2014). A, *SNVs identified and annotated by Variant Effect Predictor analysis (Ensembl). When more than one variant was present in a single gene, the most deleterious was annotated according to the colour-coded legend reported at the bottom of the figure.* ***B,*** *Blue and red bars indicate the mutation rate of OAC genes reported in Dulak et al., 2013; and Weaver et al., 2014, respectively [37].*

1.8.4 Study of the effects of mutations detected in *ELMO1*

ELMO1 and DOCK2 form the highest frequency, gain-of-function mutation "target (A protein-protein interaction pair) in OAC. As such one of these (ELMO1 became the focus for identifying functional models). Thus, at the same time of analysing the WES data of UPS, we also studied the effects of wild type *ELMO1*, and *ELMO1* with the F59L mutation detected in the OAC cases by WES by Dulak et al. [41]. We studied the effects of wild type and mutant *ELMO1* on OAC cell lines FLO-1 and OE19 growth by clonogenic assay. We have also developed a Streptavidin-Binding Peptide (SBP)-tagged affinity purification method in combination with label-free Sequential window acquisition of all theoretical mass spectra (SWATH) mass spectrometry (MS) to identify novel binding proteins for the gain-of-function mutant *ELMO1*. This identified an elevated interaction with another oncogenic protein encoded by the *AGR2* gene and validates this proteomics discovery platform to further advance function of new mutated proteins.

CHAPTER TWO
Aims and Objectives

2.1 Head and Neck

The main aims of analysing the DNA and RNA sequences of head and neck cancer patients is to set the parameters of the CLC-bio software to detect specific somatic mutations to the tumour tissues, and to detect the list of genes with somatic mutations in these patients.

We were provided with the whole exomes DNA from the tumour, normal adjacent tissue and germline blood of each of the five patients (two old and three young) of head and neck cancer. We will import these sequences files into the CLC-bio software and analyse them by comparing the sequence of the tumour to the blood and normal adjacent tissue to detect somatic mutations that are specific to the tumour tissues. We will use CLC-bio to detect areas with CNV in the tumour compared to the normal adjacent tissue and blood sequences.

We were also provided with the total RNA sequences of the tumour tissues from four of the five patients. We will compare the tumour DNA sequence with the RNA sequence to look for the expressed detected somatic mutations at the RNA level.

<u>The questions to be solved:</u>

1- What are the best CLC-bio parameters with the best sensitivity and specificity?
2- What is the best tissue to be used as a control to define somatic mutations, blood or normal adjacent tissue?
3- Are there any differences/similarities between the young and old patients in the genes with somatic mutations?
4- What is the percentage of detected somatic mutations being expressed in the RNAseq?
5- Are there any differences/similarities in CNV between the patients?

2.2 Undifferentiated pleomorphic sarcoma (UPS)

We have been provided with the fresh frozen tissues of the tumour and normal adjacent tissues of 20 UPS. We will extract the DNA from both tissues of all patients and send the DNA for whole exomes sequence by illumine. The sequences files will be imported and analysed in the CLC-bio software to detect specific somatic mutations to the tumours using the parameters set for the head and neck patients. Then we will look for common mutated genes or pathways in the UPS which can serve as therapeutic targets or biomarkers.

At the time of analysing, these sequences are analysed by another mutation calling method: MuTect. Thus, we will compare the results of each software and see which one is better in detecting real somatic mutations.

We will also extract DNA from two different parts of one of the tumours and compare their DNA sequences to the sequence of the normal adjacent tissue to see if there is any difference in the detected mutations in the two regions.

We will purify the total RNA from three UPS patients (tumours and normal adjacent tissues) and send them for sequences to compare them to the tumour DNA sequence to look for the expressed somatic mutations at the RNA level. We will also use the CLC-bio to study the different expression of genes based on the number of reads of the RNA in the tumour and normal adjacent tissues to detect genes that are overexpressed or suppressed in the tumour and study the pathways of these genes.

Protein lysate from one tumour was processed using MS to identify total proteins present in the tumour (performed as a collaboration between the Hupp laboratory and Dr Borek Vojtesek's laboratory, Brno, Czech Republic). So we will see which of the mutations detected at the genomic level are translated to protein.

The questions to be solved:

1- What is the number and type of somatic mutations detected in UPS?
2- Are there any common mutated genes or pathways?
3- Is there any difference between CLC-bio software and MuTect in detecting somatic mutations?
4- Is there any difference in mutated genes between the two different regions of the same UPS tumour?
5- What is the percentage of expressed somatic mutations which we can define by RNAseq?
6- What are the genes and pathways that have different expression profile between the normal and tumour tissues?
7- Are there any detected mutations at the genomic level translated to protein?

2.3 Investigating the effects of mutations detected in *ELMO1* gene in OAC

We would like to study the effects of wt type and mutant (F59L) *ELMO1* gene on the growth of OAC cell lines. We will study the expression of ELMO1 in some OAC cell lines such as OE19 and FLO1, and study the effects of the expression of wt and mutant ELMO1 on the growth and number of colonies of these cell lines by clonogenic assay.

We will also develop a strep-tag binding pull down method to find out wt and mutant ELMO1 binding proteins in OAC cell lines using SWATH-MS. We will then validate some of the important ELMO1 binding proteins by PLA.

Depending on time allowing, I will study some of ELMO1 interactions to other proteins and their effects on cell growth and mobility.

If time permits, these methods will be applied to investigate the effects of the detected mutated genes in UPS.

The questions to be solved:

1- What OAC cell lines express ELMO1?
2- Are there any growth differences between cells with no ELMO1 expression, with wt ELMO1 or with F59L ELMO1 expression?
3- What are ELMO1 binding proteins in OAC cell lines?
4- Are there any differences in the fold-change in the wt and mutant ELMO1-binding proteins?
5- Can these methods provide a routine approach to identify mutated protein functions in cancer?

CHAPTER THREE
Materials and Methods

3.1 Head and Neck cancer patients' DNA and RNA sequence

The DNA and RNA sequences of the five patients of head and neck cancer were provided to us by Prof. Karin Nylander (department of medical bioscience, Umea University, Sweden). For all patients, the DNA sequence was provided from the tumour, normal adjacent tissue (which is 7–10 cm distant from the tumour), and blood (except in patient 82 the sequence was derived from the tumour and blood only). From blood (buffy coat) they extracted DNA with illustra nucleon genomic DNA extraction kit according to the manufacturer manual. From tumours and normal adjacent tissues, fresh biopsies were homogenized in lysis buffer from Allprep DNA/RNA/miRNA Universal kit from Qiagen. After homogenization, extraction was continued according to the protocol in the kit. RNA sequence was provided from the tumour tissues only for all patients except 82. DNA and RNA concentrations were measured using Nanodrop. They were sent to Otogenetics for sequencing. For the tumour samples the HiSeq2500 PE100-125 kit was used to sequence the whole exomes from gDNA, paired-end 2x100-125 or PE100-125 (read length). The estimated average on-target coverage is 100x. For the normal adjacent tissues and blood, the HiSeq2000/2500 PE100 kit was used, paired-end 2x100 or PE100 (read length). The estimated average on-target coverage is 30x. HiSeq2500 PE100 sequencing used to sequence the whole RNA.

3.2 Sarcoma DNA and RNA sequences to detect somatic mutations

3.2.1 Purification of DNA and RNA

DNA was extracted from the 20 sarcoma tumours and their matched normal tissues using the ChargeSwitch gDNA Mini Tissue kit from Invitrogen following the manufacturer's protocol.

RNA were extracted from normal and tumour tissues of patients 55, 66 and 73 using RNeasy RNA extraction kit from Qiagen following the manufacturer's protocol.

3.2.2 Sequencing of DNA

Exome Sequencing was performed using Agilent V5+UTR Exome Capture Kit (75Mb); Illumina, 100bp paired-end reads using a coverage of Tumour: Normal pairs (100X/30X). Paired de-multiplexed fastq files were generated using CASAVA software (Illumina) and initial quality control was performed using FastQC.

3.2.3 Sequencing of RNA

RNA sequencing was performed using Illumina HiSeq2500 100bp paired-end reads.

3.2.4 Read mapping of DNA reads and variant detection

Paired de-multiplexed fastq files from DNA-exome libraries were imported into the CLC Biomedical Genomics Workbench (version 2.5). Adaptor sequences and bases with low quality were trimmed and reads were mapped to human genome 19 (HG19). The *remove duplicate mapped reads* tool was used to remove paired reads that have the same start and end coordinates and are, thus, probably PCR duplicates. Variants were detected in the exome data with the CLC Bio Probabilistic Variant Caller using the following parameters: Minimum coverage (number of reads) = 5; Minimum frequency = 5%; Minimum number of variants = 2; Variants in normal germline DNA=0, and the coverage in the germline DNA should be at least 5 reads at the variant site. The CLC workflow of variants detection in the tumour is shown in figure 3.1.

3.2.5 Read mapping of RNAseq

Paired de-multiplexed fastq files from RNAseq libraries were trimmed for stretches of adapter sequences, joined into a single read followed by quality trimming using commands from the CLC Assembly Cell. This resulted in three different fastq files: paired-end data, single-end data after joining of paired reads, and single-end data after elimination of one of the paired reads due to e.g. low quality. All three types of fastq files were then imported batchwise into the CLC Biomedical Genomics Workbench (version 2.5).

3.2.6 Detection of expressed somatic mutations in the RNAseq

In order to detect the expressed somatic mutations in the RNA, the tumour RNAseq was compared to the tumour DNAseq of each patient (55, 66 and 73). The minimum coverage and number of variants were set to 1 in the RNA reads, while in the DNA the parameters were not changed and left as 5 for minimum coverage and 2 for the minimum number of variants. The workflow for detection of variants in both RNA and DNA is shown in figure 3.2.

In order to get expressed somatic mutations; The tool remove germline variants was used to remove the germline variants from the variants found in both tumour DNA and tumour RNA.

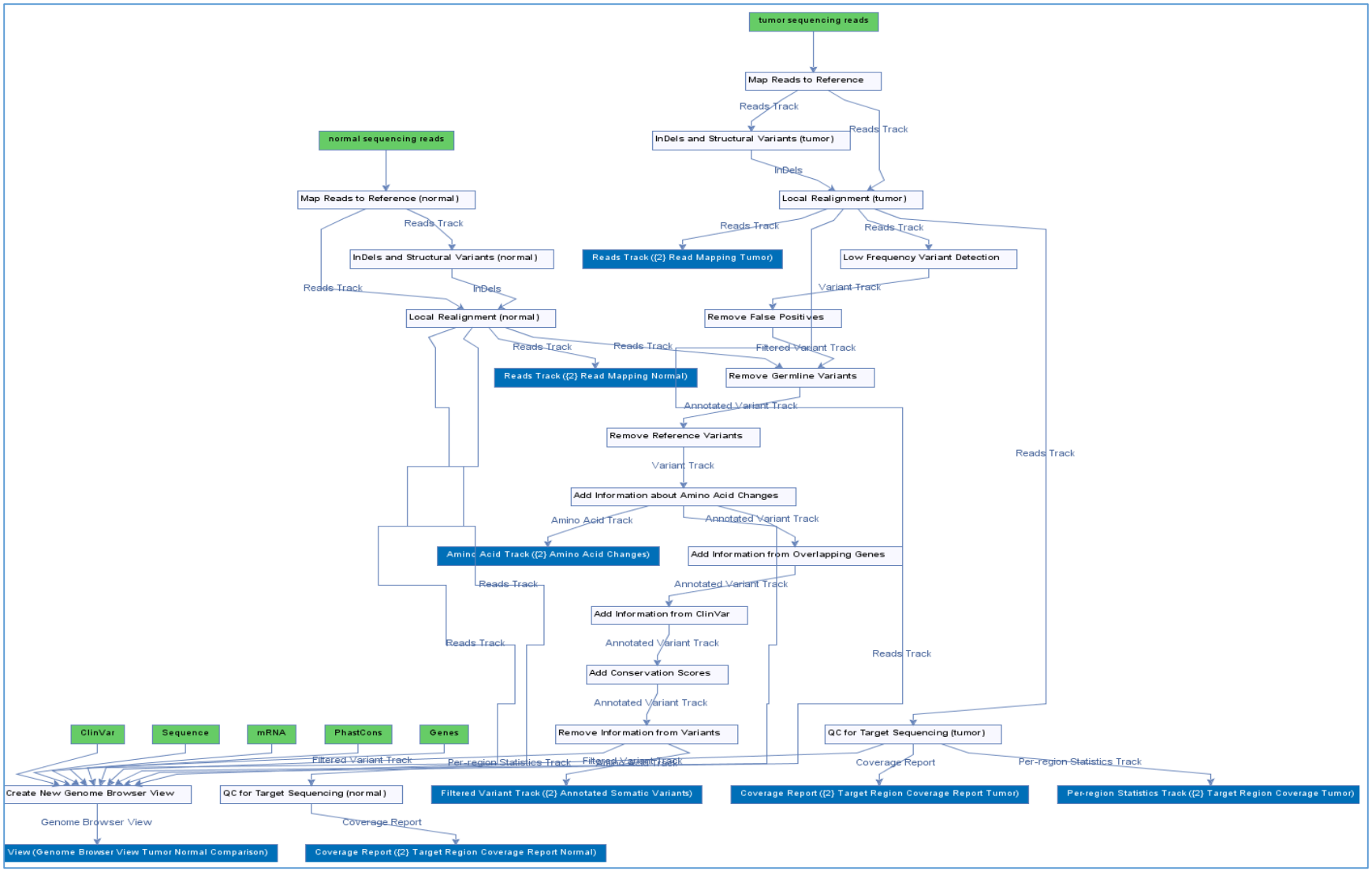


Figure 3.1. Workflow of the identification of somatic mutations in the tumour DNA sequence by comparing the tumour sequences to the sequence of the normal tissue of the same patient in CLC Biomedical Genomic workbench.

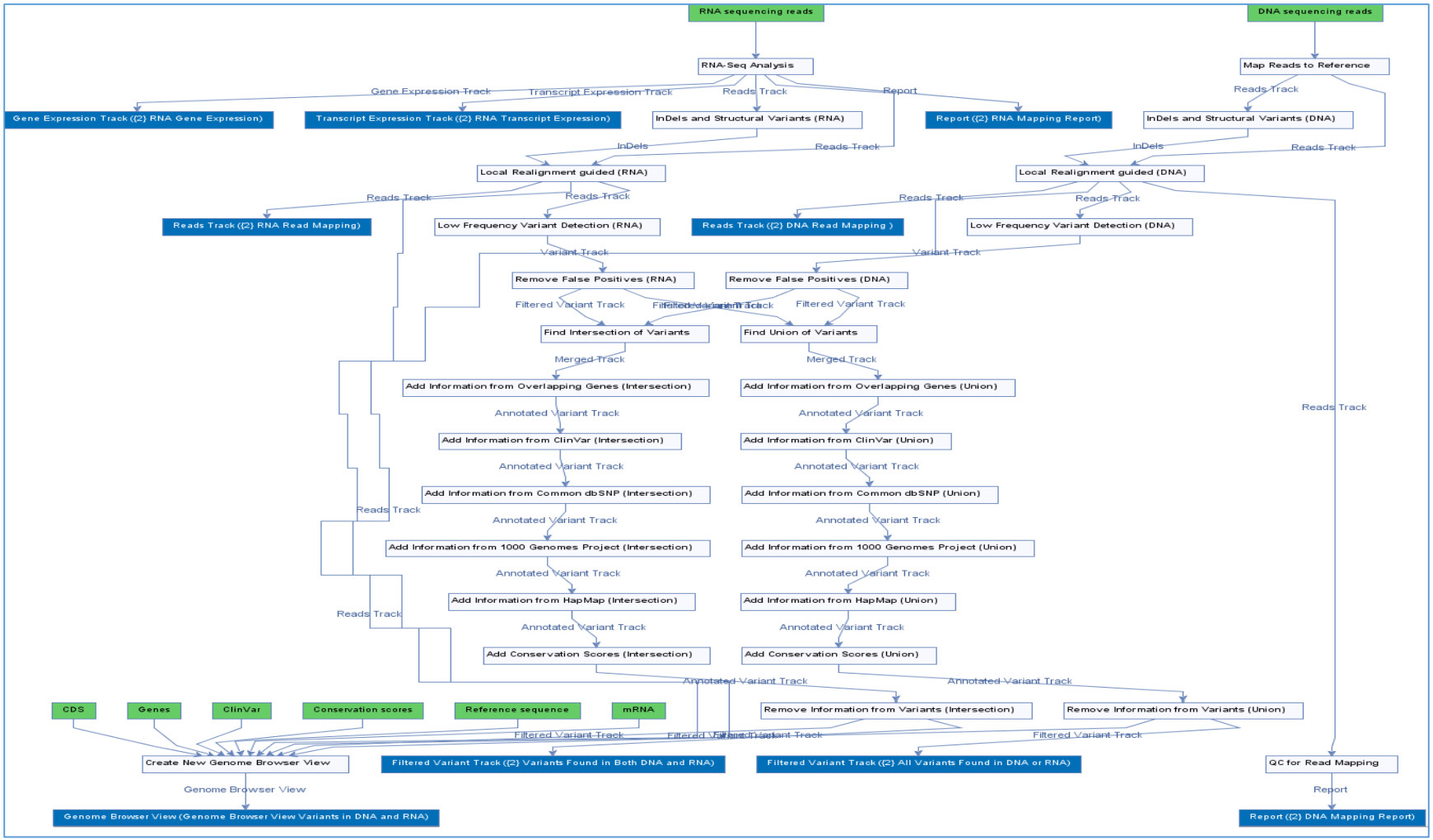


Figure 3.2. The workflow of identification of variants found in the tumour DNA and tumour RNA in CLC Biomedical Genomic workbench.

3.3 General microbiological techniques

All the microbiological techniques were carried out using sterile glass, plastic apparatus and under strict aseptic conditions. Different techniques used in the lab have been described below:

3.3.1 Transformation of bacterial competent cells

Between 50 to 200 ng of plasmid DNA or 5 µl of ligation product was added to a 50 µl cell aliquot of *E.coli* DH5α cells for plasmid propagation. The cell aliquots containing DNA were incubated on ice for 30 mins, placed in a water bath at 42^{o}C for 45 seconds and back on ice for 2 mins. Under sterile conditions, 500 µl of fresh LB medium was added to the cell aliquot and incubated shaking at 37^{o}C for 60 mins. LB agar 10 cm dishes were prepared (LB containing 1.5% (w/v) agar) with 100 µg/ml Ampicillin (#A9618, Sigma) or 25 µg/ml Kanamycin (#11815032, Invitrogen), depending on the appropriate antibiotic resistance encoded by plasmid. All the unused plates were stored at 4^{o}C upside down, sealed in parafilm and pre-warmed in 37^{o}C incubator before use. In a sterilized area, 50-200 µl of the culture was plated out onto the pre warmed LB plate using the glass spreader. Plates were incubated overnight upside down in 37^{o}C incubator.

3.3.2 Purification of plasmid DNA

Under sterile conditions, a single colony was picked from the transformed plate using a sterile tip. The colony was dropped into a 5 ml LB medium (100 µg/ml Ampicillin or 25 µg/ml Kanamycin) in a 15 ml falcon tube (for small scale purification). The tube was incubated overnight at 37^{0}C, 220 rpm and the following day the tube was centrifuged at 4000 rpm for 10 mins at 40C. Following centrifugation, the tip and supernatant was discarded carefully. QIAprep Spin Miniprep Kit (#27106, Qiagen) instructions were followed to extract and purify plasmid DNA which was then further eluted in 30-50 µl of elusion buffer. The concentration of the DNA was measured on a Nanodrop and stored at -20^{o}C.

For a larger scale purification of plasmid DNA, colony was dropped in a similar way but the volume of LB containing appropriate antibiotic was increased to 150-200 mL and grown overnight at 37^{o}C. HiSpeed Plasmid Kit (#12663, Qiagen) was used for extraction and purification of plasmid DNA, and finally DNA was eluted in 1 ml of elusion buffer and stored at -20^{o}C.

3.4 Molecular biology techniques

All primers were designed and ordered from Sigma Genosys, whereas all enzymes and buffers were ordered from New England Biolabs unless stated otherwise.

The plasmids used and their sources were: pEGFP-C1 and pEXPR-IBA105 (Addgene)

3.4.1 DNA quantification

The concentration of DNA was measured using Nanodrop spectrophotometer (absorbance at 260 nm).

3.4.2 Polymerase chain reaction

A suitable template as cDNA or plasmid DNA is used to amplify the desired gene by Polymerase chain reaction (PCR). The primers were designed in such a way that forward primer contains a restriction site which cuts at the 5'end of the gene and a reverse primer which has a site attached at the 3' end of the primer.

While designing the primers, it was ensured that the gene was in frame with the N-terminal of the vector sequence. Few base pairs were added to the 5'end of the primers before the restriction sites in order to allow efficient binding of the restriction sites to the primer and allow significant digestion of the insert. Also, primer length was manipulated to get a suitable melting temperature and presence of GC at the primer extremities was favoured. PCR reactions were carried out on ice in nuclease free tubes with 2 X high Fidelity Phusion Master mix (#M2075 NEB), 100ng of plasmid DNA, 0.5 µM of each primer, nuclease free H2O made up to 50 µl.

The PCR program was as the following:

- incubate at 95 C for 2 minutes
- incubate at 95C for 20 seconds
- incubate at 58C for 40 seconds
- incubate at 72C for 1 minutes
- cycle to step 2 for 35 cycles
- incubate at 72C for 5 minutes
- 4C forever

5-10 µl of PCR reaction was run on a 1.5% (w/v) agarose gel to check if the desired band for the amplification of the gene was obtained. Multiple PCR reactions were run on 1.5% (w/v) agarose gel until maximum separation of the desired and non-specific bands was achieved. The desired band was cut using a sterile scalpel and this excised gel was purified by using Gel

extraction kit (#28704, Qiagen) manufacturer's protocol. The DNA was eluted in 30-50 µl of elusion buffer (Buffer EB) and the concentration of the DNA was determined using Nanodrop (OD600nm).

3.4.3 *ELMO1* primers

The primers used to amplify ELMO1 cDNA are shown in table 3.1.

	Forward 5'to 3' primer with *EcoR1* restriction site	**Reverse 3' to 5' primer with *BamH1* restriction site**
ELMO1 primers to be cloned in Pexpr-iba105 vector	GGGGGAATTCGATGCCGCCACCCG CGGACATCGTC	CCCCGGATCCTCAGTTACAGTCATA GACGAAGTC
ELMO1 primers to be cloned in EGFP-C1 vector	ACCGGAATTCTATGCCGCCACCCG CGGAC	CGGTGGATCCTTAGTTACAGTCAT AGACGAA

Table 3.1. primers to amplify whole ELMO1 cDNA*. Primers were designed using primer 3 website. Forward and reverse primers have EcoR1 and BamH1 restriction sites (underlined) respectively to be ligated in Pexpr-iba105 and EGFP-C1 vectors. Additional one base (green) was added to each of the forward primer to make the expression of ELMO1 inframe*

3.4.4 Restriction digestion of the purified PCR product and destination Vector DNA

For these above purified PCR products, double digestion was performed using two different enzymes and compatible buffers recommended by supplier. The destination vector along with the insert was digested using these two enzymes and incubated for 1-2 hours depending upon the amount of the vector and insert. To ensure efficient digestion has taken place, controls were set up such as single digestion using enzyme 1 and enzyme 2 individually as well as no enzymes.

Restriction Enzyme Recognition sequence 5'-3' (/ indicates cut site):

- BamH1 G/GATCC
- EcoRI G/AATTC

3.4.5 Ligation of vector and insert

Ligation of vector and insert was carried out using T4 DNA ligase (Promega), following manufacturer's instructions. A standard amount of vector (100 ng) was used and the amount of insert required was calculated using the following formula:

$$\text{Insert ng} = \frac{\text{Vector ng} \times \text{Insert size kb} \times \text{Molar ratio of insert}}{\text{Vector size kb vector}}$$

A 1:1 and 3:1 molar ratio of insert to vector was always tested.

Ligation reactions were as follows:

- 1 µl ligase buffer (10x)
- 100 ng vector
- *X* ng insert
- 1 µl T4 DNA ligase
- Nuclease free water to 10 µl water

Ligation reactions were carried out at room temperature for 1 hour. Following which 2.5 µl of the mix was transformed into DH5α competent cells and plated onto Ampicillin or Kanamycin containing LB agar plates. Colonies were selected and plasmid DNA obtained using the Qiagen Mini-prep kit.

3.4.6 Site directed mutagenesis

Site directed mutagenesis was carried out using methods from Stratagene. The primers used to make F59L mutation (TTC > TTA) in ELMO1 were TGCCGATAGTTCAAACTTATATATCACAGA and its reverse complement TCTGTGATATATAAGTTTGAACTATCGGCA.

The PCR reaction was set up in nuclease free tubes using 2 x Pfu master mix (Rovalab) as follows:

- 25 µl 2x Pfu master mix
- 5 µl Band doctor (Rovalab)
- 20 ng template DNA
- 1 µl forward primer (20 µM stock)
- 1 µl reverse primer (20 µM stock)
- Nuclease free water up to 50 µl

Thermal cycling conditions:

- Incubate at 95°C for 2 minutes
- Incubate at 95°C for 1 minute
- Incubate at 60°C for 1 minute
- Incubate at 72°C for 5 minutes
- Cycle to step 2 for 16 cycles
- Incubate at 72°C for 5 minutes
- Hold at 4°C forever

Post PCR 1 µl DpnI (20000 U/ml, NEB) was added directly to each reaction tube and incubated at 37°C for 1 hour. Dpn1 was inactivated after this time by incubating at 65°C for 10 minutes.

PCR reactions (2.5 µl) were transformed into DH5α competent cells and plated onto Ampicillin or Kanamycin containing LB agar plates. Colonies were selected and plasmid DNA obtained using the Qiagen Mini-prep kit.

3.4.7 Agarose gel electrophoresis of DNA

Agarose gel electrophoresis was used to separate, identify and purify DNA fragments.

Agarose gels (1 %) were prepared by dissolving agarose in 1xTAE. To detect the DNA the SYBR®Safe (Invitrogen) was added to the gel prior to pouring the gel. Loading dye (6x) was added to the DNA samples prior to loading onto the agarose gel; the gel was run in 1x TAE buffer at 100 V for~1 hour.

1x TAE buffer

- 40 mM Tris
- 1 mM EDTA
- pH 8

6x DNA loading dye

0.25 % (w/v) bromophenol blue
0.25 % (w/v) xylene cyanol
15 % (v/v) glycerol

3.4.8 Reverse transcription of total RNA to cDNA

Omniscript Reverse transcription kit (Qiagen) was used in order to make cDNA from RNA. The preparation of reaction mixture is shown in table 3.2.

Component	Volume/ reaction	Final concentration
Master mix 10X Buffer RT	2 µl	1 x
dNTP mix (5 mM each dNTP)	2 µl	0.5 mM each dNTP
Oligo –dT primer (10 µM)	2 µl	1 µM
RNAse inhibitor (10 units/ µl)	1 µl	10 units (per 20 µl)
Omniscript Reverse transcription	1 µl	4 units (per 20 µl)
RNAse free water	variable	
Template RNA Template RNA, added at step 5	variable	Up to 2 µg (per 20 µl reaction)
Total volume 20 µl	Total volume 20 µl	

Table 3.2. preparation of reaction mix for making cDNA from RNA.

3.4.9 DNA sequencing

Sequencing was carried out by Source Bioscience (LifeSciences). Plasmid DNA was obtained from clones by using the Qiagen Mini-prep kit, following manufacturer's instructions. Plasmids were sequenced using stock primers from source bioscience.

3.5 Validation of detected mutations

3.5.1 Validation of variants detected in *DMKN* gene in head and neck cancer patients

The primers to amplify the region of *DMKN* gene with the variants were designed using primer 3 website. The primers are 5' GCCCGTCTTGGTGCTCTGTCT and 3' CCCGGCAGGGACCACTAGGGCC. The gene was amplified and tested for the variants in the department of medical bioscience, Umea University, Sweden.

3.5.2 Validation of variants detected in sarcoma patients

In order to validate some of the detected variants; primers have been designed to amplify the regions around the variant (table 3.3) and sent for Sanger sequence.

Gene name	Patient	5'primer	3'primer
FAT3	73	CCAGAGACAAAGACTTAGGTTCT	TGACAGAGAGACAGGCCGCCCTT
IL11RA	55	AGGAGGAGTCCATCCACAGG	ATGCTGTACGAGTCAGTGCC
SEMA6A	55	CAAGCCGTGGATTACGGAGATTA	AGCGGGTGCGTCTTGATGAAGTT
MTCH2	55	TGACCACGTTATCAAGGAG	AGAAGGCGAGGAACAAGAC

Table 3.3. primers designed to amplify regions around the detected variants in FAT3, IL11RA, SEMA6A and MTCH2 genes.

3.5.3 RNA editing variants in sarcoma patient 55

Primers were designed around the variant in each of the 10 genes selected for RNA editing validation in patient 55 (table 3.4). The black bases are the primers which are complementary to the gene sequence to be amplified, and the red bases are the adapter sequence linked to the primer. The primers were used to amplify the genes directly from the RNA of patient 55 using the superscript one-step RT-PCR (Invitrogen 12574018) according to the manufacturer manual. The amplified amplicons were sent for deep sequencing which was performed by Peter Muller (Borek Vojtesek laboratory, Brno, Czech Republic).

Chromosome	Region	Reference	Allele	Tumour frequency	Gene name	5' to 3' primer	3' to 5' primer
6	33289002	A	G	33.33	*DAXX*	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGACCCCACAAATGCTGAAAACAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGCCGGATCTCTGCCACATAG
11	116633907	T	A	28.57	*BUD13*	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGAAGACCTACCCTCAAACAGACA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGTGTCATGACGGTCCTTCCTA
16	30958168	T	C	25	*FBXL19*	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCGTCTCCTGCTGCGTCAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGGCCGTGAGGAGGTGAAC
21	30341916	T	A	25	*LTN1*	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTCCTCACGTCTCTAGTTGCTGGG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGATATTACTGCCGAGGACT
X	107084433	C	T	23.33	*MID2*	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCCACGCACCCCAACAA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATTCTCATGGTCCAGGCAGGTG
2	169622111	T	A	22.22	*CERS6*	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTATGTTTGCCGTGGTTTTTATCA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGTAAGGTCCAACGATCTCCCAGC
15	41029488	T	A	22.22	*RMDN3*	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTTTTCACTGACTAATAGGACTC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGATAGGTGATAAGAAAGCAATGT
18	8636316	C	A	22.22	*RAB12*	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGTGAAGCAAGTGCCAAGGATAAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGCATCTTTTTCAGAATGTCATCG
7	100730901	T	C	20	*TRIM56*	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCCGGGCCTGTGGAGAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCCCCGGTGCTGGTGCC
12	49420202	G	A	20	*KMT2D*	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTTTTCACTGACTAATAGGACTC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCAGCAGCTGTCCGATGG

Table 3.4. primers designed to amplify regions around RNA edits in 10 genes of sarcoma patient 55*. Adapter sequence (red) was attached to the primer sequence.*

3.5.3.1 Validating the edit in *MAP3K5* gene

The primers were designed to amplify and validate the RNA edit in MAP4K5 gene transcripts as shown in fig 3.3 below.

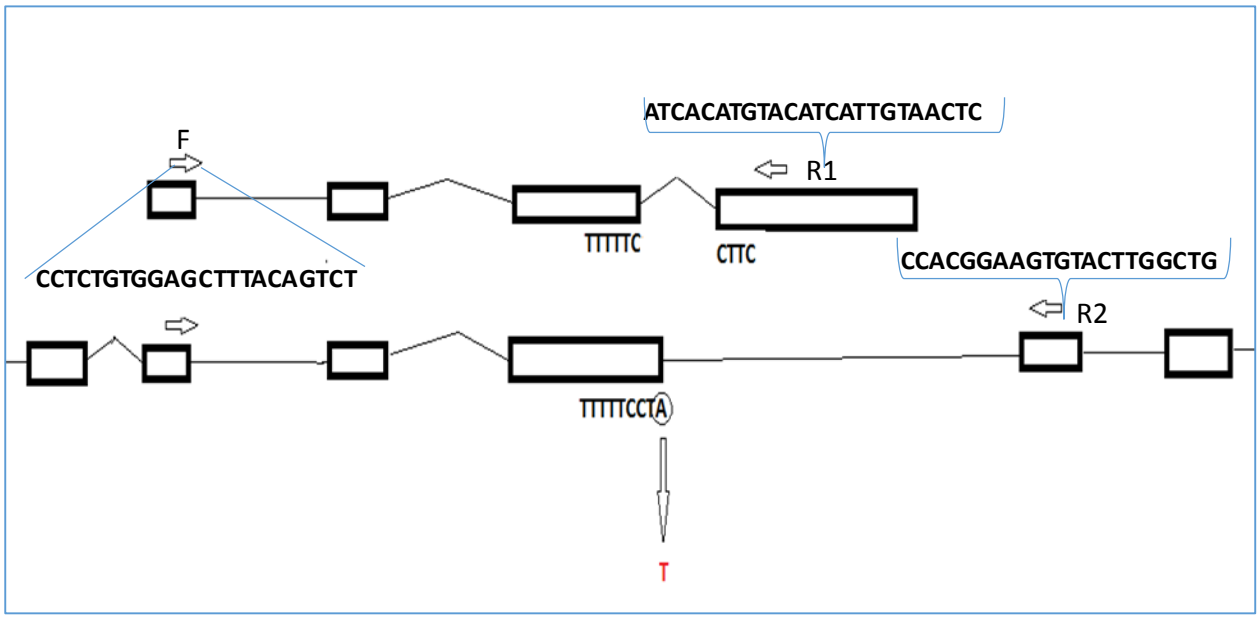


Figure 3.3. the primers locations and sequences for the MAP4K5 gene transcripts to validate the RNA edit A>T.

3.6 Cell lines

3.6.1 Maintenance of cell lines

Cell lines were maintained in the appropriate medium supplemented with 10% (v/v) foetal bovine serum (FBS; Autogen Bioclear). Cells were incubated at 37°C in a humidified 5% CO2 incubator or 10% CO2 for specific cell lines (table 3.5).

Cell line	Cell Line Origin	Culture medium	Origin
FLO-1	Distal oesophageal adenocarcinoma	DMEM (Gibco, Invitrogen)	Prof. Ted Hupp
OE19	Human Caucasian oesophageal carcinoma	RPMI 1640 (Gibco, Invitrogen)	Prof. Ted Hupp
OE33	Human Caucasian oesophageal carcinoma	RPMI 1640 (Gibco, Invitrogen)	Prof. Ted Hupp

Table 3.5. oesophageal adenocarcinoma cell lines.

Cells were maintained in 10 cm tissue culture dishes and were sub-cultured 2-3 times a week once the confluency rate reached ~90%. The appropriate media needed for the cell line was warmed up to 37°C in the water bath along with trypsin. Medium was removed and cells were

washed with 7-8 ml of sterile PBS. The PBS was poured off and cells were trypsinized using 2 ml of trypsin –EDTA 0.5% (#25300, Gibco, Invitrogen) with an incubation period of 5 minutes at 37°C. Following 5 minutes of trypsinisation, 8 ml of pre-warmed fresh media was added and 2 ml was transferred to a new 10 cm dish containing 8 ml of fresh relevant media. The dishes were then incubated at 37°C overnight.

3.6.2 Transfection of mammalian cells

Cells were seeded in such a way that they reached a confluency of 50-70% in order to transfect. Transfection was carried out under laminar air flow in tissue culture room. For transfection, sterile eppendorfs as well as filter sterile tips were used after spraying with 70% ethanol. For preparing transfection reaction mix, 100 µl of respective serum free media (based on the cell line) was taken in a sterile eppendorf. To this media, appropriate amount of DNA was added (0.5, 1 and 5 µg) along with 5 µl of transfection reagent. The transfection reagent used was attractine (Qiagen). The reactions were then incubated for 15 minutes. Subsequently, the transfection reactions were added to the cells that were seeded before 12-14 hours previously. The plates were incubated for 18- 24 hours in a 37°C with 5% CO2.

3.6.3 Harvesting of cells

The cells were washed once within the 10cm plate with PBS and 1ml of PBS was added in order to harvest the cells. The cells were harvested by using a plastic scraper and harvested cells were then transferred into a fresh sterile eppendorf. The cells were centrifuged at 5,000 rpm for 5 minutes and the pellet was snap frozen and stored in -80^{o}C.

3.6.4 Lysis of cells

Lysis buffers:

- Urea lysis buffer	- 0.5% NP-40 lysis buffer
6.24M urea	25mM Tris-HCl, pH 7.4
0.1M DTT	150mM NaCl
0.05% Triton X-100	1mM EDTA
25mM NaCl	0.5% NP-40
20mM HEPES-KOH, pH 7.6	5% glycerol
1X Protease Inhibitor mix	1X Protease Inhibitor mix

The cells were incubated on ice for 30 minutes for lysis and centrifuged at 10000 rpm and the supernatant were transferred to a sterile eppendorf. The protein concentration within the lysate was determined by Bradford assay (Bio-Rad) after reading the plate at wavelength of

595 nm. The samples were then prepared using 2X sample buffer containing 1 M DTT and normalized with water. The prepared samples were heated at 85^{o}C for 5 minutes and loaded onto an appropriate percentage SDS gel.

3.7 SDS gel preparation and Immunoblotting

3.7.1 2X Sample buffer preparation

- 2 ml Tris (1 M, pH 6.8)
- 4.6 ml glycerol (50%)
- 1.6 ml SDS (10%)
- 0.4 ml bromophenol blue (0.5%)
- 0.4 ml β- mercaptoethanol

The samples after normalizing were made up to a certain amount with 2X sample buffer. Based on the size of protein, the lysate/ protein was run on an appropriate percentage of SDS polyacrylamide gel. The samples were loaded onto the gel along with a pre-stained ladder and the gel was run in 1X running buffer at 150 V for 1 hour at RT. Solutions needed for Tris glycine SDS - polyacrylamide gel electrophoresis- three different size of combs: 0.75 mm, 1.0 mm and 1.5 mm (10, 15 wells), the appropriate percentage of SDS gel was made based on the molecular weight of the protein. The ratio of all the ingredients has been mentioned below (table 3.6):

Solution Components	Component volumes (ml) per gel mold volume of							
	5ml	10ml	15ml	20ml	25ml	30ml	40ml	50ml
6%								
H2O	2.6	5.3	7.9	10.6	13.2	15.9	21.2	26.5
30% Acrylamide	1.0	2.0	3.0	4.0	5.0	6.0	8.0	10.0
1.5M Tris (pH8.8)	1.3	2.5	3.8	5.0	6.3	7.5	10.0	12.5
10% SDS	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
10% ammonium persulfate	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
TEMED	0.004	0.008	0.012	0.016	0.02	0.024	0.032	0.04
8%								
H2O	2.3	4.6	6.9	9.3	11.5	13.9	18.5	23.2
30% Acrylamide	1.3	2.7	4.0	5.3	6.7	8.0	10.7	13.3
1.5M Tris (pH8.8)	1.3	2.5	3.8	5.0	6.3	7.5	10.0	12.5
10% SDS	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
10% ammonium persulfate	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
TEMED TEMED	0.003	0.006	0.009	0.012	0.015	0.018	0.024	0.03
10%								
H2O	1.9	4.0	5.9	7.9	9.9	11.9	15.9	19.8
30% Acrylamide	1.7	3.3	5.0	6.7	8.3	10.0	13.3	16.7
1.5M Tris (pH8.8)	1.3	2.5	3.8	5.0	6.3	7.5	10.0	12.5
10% SDS	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
10% ammonium persulfate	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
TEMED	0.002	0.004	0.006	0.008	0.01	0.012	0.016	0.02
12%								
H2O	1.6	3.3	4.9	6.6	8.2	9.9	13.2	16.5
30% Acrylamide	2.0	4.0	6.0	8.0	10.0	12.0	16.0	20.0
1.5M Tris (pH8.8)	1.3	2.5	3.8	5.0	6.3	7.5	10.0	12.5
10% SDS	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
10% ammonium persulfate	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
TEMED	0.002	0.004	0.006	0.008	0.01	0.012	0.016	0.02
15%								
H2O	1.1	2.3	3.4	4.6	5.7	6.9	9.2	11.5
30% Acrylamide	2.5	5.0	7.5	10.0	12.5	15.0	20.0	25.0
1.5M Tris (pH8.8)	1.3	2.5	3.8	5.0	6.3	7.5	10.0	12.5
10% SDS	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
10% ammonium persulfate	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
TEMED	0.002	0.004	0.006	0.008	0.01	0.012	0.016	0.02

Table 3.6. SDS gel ingredients: *different percentage of gel based on thickness to run higher and lower molecular weight proteins.*

Stacking gel:
30 % acrylamide mix 5 % (v/v)
1 M TRIS (pH 6.8) 0.13 M
10 % (w/v) SDS 0.1 % (v/v)
10 % (w/v) APS 0.1 % (v/v)
TEMED 0.1 % (v//v)
H2O to final volume

3.7.2 Western blot

Once the gel was run, it was transferred onto nitrocellulose membrane in such a way: Black side of cassette – sponge - blotting paper – gel - membrane – blotting paper –sponge. 10 X transfer buffer was diluted down to 1 X with an addition of 200 ml of methanol to enhance transfer of small proteins. The western blot was performed at 100 V for 1 hour with an ice block. Subsequently, the membrane was ink stained and blocked with 5% (w/v) skimmed milk in PBS-T (0.1%) for 1 hour. Following on, primary antibody was diluted in 5% milk (in PBS-T) and added to the blocked membrane and incubated for 1 hour at RT or overnight at 4°C. A list of the primary antibodies used is shown in table 3.7. The blot was then washed three times with PBS - T (0.1% (v/v)) and HRP conjugated appropriate secondary antibody was diluted in blocking buffer and added to the blot and again incubated for hour. The blots were then washed three times again with PBS-T 0.1% (v/v) and the results were analysed by overlaying the blots with ECL (1+2) for 2 minutes.

ECL I	**ECL II**
100 mM Tris HCL pH 8.5	100 mM Tris HCL pH 8.5
2.5 mM Luminol	0.02% (v/v) H2O2
0.4 mM p- Coumaric acid	-
Make up with H2O upto 100 ml	Make up with H2O upto 100 ml

Protein	Primary used	Company	Approximate molecular weight of the band KDa
ELMO1	Goat polyclonal	Abcam ab2239	~ 80
GFP	Rabbit polyclonal	Abcam ab6556	27
AGR2	Rabbit polyclonal K47	Moravian Biotechnology	18-21
B-actin	Rabbit polyclonal	Abcam ab8227	47

Table 3.7. A list of primary antibodies used to detect ELMO1, GFP, AGR2 and B-actin proteins.

3.7.3 Silver staining

Silver staining was carried out according to instruction in Pierce™ Silver Stain Kit (Thermo Scientific 24612).

3.8 SBP-tagged pull down experiment

FLO-1 and OE19 cells were cultured in 10 cm plates and transfected with 5 µg of the Strep-tag vector (pEXPR-IBA105), empty, wt ELMO1 and F59L ELMO1 using attractine. After 24h of incubation in 37°C incubator, cells were lysed with 500 µl of 0.5NP40 lysis buffer. Then 15 µl of Streptavidin agarose beads were added to the lysate and incubated in the shaker at 4°C for 2 hours. The beads were then washed 3X with 0.5NP40 buffer and ran on SDS-PAGE. Silver staining, and immunoblotting with ELMO1 specific antibodies were carried out. The beads were then sent to Borek Vojtesek laboratory, Brno, Czech Republic, to identify ELMO1 binding proteins by SWATH-MS.

3.9 SWATH-MS

SWATH method for label-free quantification of proteins in complex mixtures was set-up according to previously published methods [43]. TripleTOF 5600+ (AB-SCIEX, Toronto, Canada) operated in high sensitivity positive mode. Random precursor ion peaks were extracted from TOF-MS and the approximate chromatographic peak width was defined to correctly establish SWATH method so that at least ten data points were acquired across a peak. Four randomly extracted precursor peaks from TOF-MS were evaluated and the peak width at FWHM was in average 1.5 min, so the cycle time of SWATH was set to 3.5 s. With the defined cycle time an optimal SWATH width of 20 Da with 1 Da overlap was calculated, with accumulation time 98 ms per SWATH. Precursor range was selected from 400 amu up to 1100 amu. Product ion range was scanned from 300 amu up to 1600 amu and rolling collision energy was used with collision energy spread (CES) of 10 mV. Spectral library for SWATH data mining was measured from 1 _L pool of cell lysates (approx. 1 _g/_L protein concentration). Mass spectrometer TripleTOF 5600+ (AB-SCIEX, Toronto, Canada) operated in data-dependent mode. During each cycle, mass spectrometer fragmented the top 20 intense precursor ions with exclusion time set to 12 s. Minimum precursor ion intensity was set to 50 cps, 100 ms accumulation time was used and 150 ms accumulation time for TOF-MS scan. For building up of spectral library 1632 proteins FDR 1% were used after Protein Pilot 4.5 (AB-SCIEX, Toronto, Canada) search using

Uniprot 2013_12 database. Spectral library was built in Peakview software 1.2.0.3 (AB-SCIEX, Toronto, Canada), only the identifications below FDR of 1% were indexed. Quantitative data (peak areas) corresponding to each protein included in spectral library were extracted from SWATH data using manual analysis in Peak view 1.2.0.3 (AB-SCIEX, Toronto, Canada).

Data were extracted using retention time window of 3.5 min, which was determined by extracting random peaks across LC gradient. Retention time window describes the LC retention time shifts between SWATH technical replicates and data dependent acquisition (DDA) measurement and specifies in which scope of retention times software should look for peaks included in spectral library (DDA measurement result). Eight peptides per protein and five product ions per each peptide were used. Extracted quantitative data were further analysed in Marker view where T-testing was done on quantitative data from all replicates originating from compared sample pair. As a result, for all proteins in spectral library protein fold changes and *p* values between chosen sample pair were calculated and are valid only for the concrete pair comparison.

3.10 Proximity ligation assay (PLA)

As means of developing new assay to demonstrate the binding between ELMO1 and AGR2, we performed PLA onto FLO-1 cells. FLO-1 cells were grown on a 19 mm cover slip for 24 hours, then they were transfected with ELMO1 (wt and mutant) with and without AGR2. In the following day coverslips were made ready for PLA to be carried out.

The cells were fixed onto slides with 4% paraformaldehyde in PBS for 20 min at RT, permeabilized for 10min in 0.25% Triton x-100 in PBS and blocked with 3% BSA in PBS for 30 min. Antibodies from different species were then incubated on the slides, goat polyclonal ELMO1 antibody (Abcam ab2239) and rabbit polyclonal AGR2 K47 antibody (Moravian Biotechnology), at a 1:250 dilution for 1 hour at RT. Following PBS washes IF coverslips were incubated with donkey anti-goat and donkey anti-rabbit antibodies for 1 hour at RT. IF coverslips were further washed in PBS stained with DAPI and mounted onto slides with fluorescent mounting medium. Proximity ligation assay (PLA) was carried out with the OLIGO duolink.

Designated protocol using anti-goat and anti-rabbit probes (Sigma; The duolink probe product numbers are 92002 (rabbit plus), 92006 (goat minus), and the duolink green detection is

92014.) The PLA coverslips were stained with DAPI and mounted in the same fashion as the IF coverslips. Images were taken at 20X using an Olympus BX51 microscope.

3.11 Clonogenic assay

FLO-1 cells were cultured in 6-well plates, and when they were 70-80% confluent they were transfected with 1 µg of pEGFP-C1 and pEXPR-IBA105 plasmids; empty, with wt ELMO1, with F59L ELMO1 using Attractine as a tranfection reagent. After 24 h, cells were trypsinized and recultured in 10 cm plates with DMEM media containing G418 (Geneticin) drug at concentration of 400 µg/ml. Cells were incubated at 37°C incubator for around one month, changing media with drug selection every week. After the one month, media were discarded and cells were washed with PBS and then stained with Leishman stain (*sigma Aldrich*) for around 15 minutes and then washed with water and allowed to dry at room temperature. The colonies were counted in each plate.

CHAPTER FOUR
Analysing somatic mutations in the whole exome sequence and RNA sequence of five patients with Head and Neck cancer

4.1 Introduction

4.1.1 Developing a cancer tissue model for applying novel genomic DNA variant-calling software to identify mutations in human cancers

The vast majority of studies that have been published for identifying mutations in DNA sequencing files from NGS use a version of MuTect, usually modified by bespoke code. This is not accessible to the vast majority of scientists. In parallel, and similar to next generation DNA sequencing technologies which have largely been commercialized, software applications with browser interfaces have also been developed by commercial groups. Thus, in a similar way to Illumina being the first to develop protocols for DNA sequencing, CLCbio have generated applications for interrogating DNA and RNAseq files. My lab has had access to beta versions of these programs and was one of the first to use them for both DNAseq and RNAseq analysis. This thesis presents how we developed and used this software and what we have learned from our analyses.

Upon starting to develop our analysis of novel software for defining cancer-specific mutations, we focused on a selected set of cancer samples that are relatively homogeneous clinically, and from which we can obtain normal adjacent tissue as well as blood for defining truly "non-tumour" DNA sequences. In addition, collaborators in the pathology field can guarantee follow-on validation. Most NGS studies use either only germline (blood) or normal adjacent tissue as the source of "non-tumour DNA". This approach is relatively controversial; as normal adjacent tissue might have "field" mutations. Thus, these papers do not address true mutations that are different from germline (if normal adjacent tissue is used), or they do not define true tumour mutations (if normal blood is used) as normal adjacent tissue might have

seed mutations not linked to the cancer itself. This is especially important due to the concept of cancer field cancerization. Head and neck cancer (HNC) was our chosen tissue to begin our analysis. The clinical group chosen is also emerging as a new focus area, so the clinical questions are novel.

4.1.2 Head and Neck Cancer (HNC)

HNC is a heterogeneous disease that can involve multiple sites within the head and neck region, such as the paranasal sinuses, nasal cavity, oral cavity, pharynx, salivary glands and the thyroid [44]. HNC remains a major medical problem with high morbidity, mortality and quality of life issues with a 5-year survival of less than 50%. Patients are often diagnosed at advanced stages with serious lymph node metastasis. The HNC is the fifth most common cancer type in the world, with approximately 550,000 new cases of HNC diagnosed annually worldwide [45]. Tumours of the head and neck have different incidences, mortalities and prognoses in different parts of the world; for example, in India, it is the most common cancer, accounting for 40% of all malignancies [46]. Head and neck squamous cell carcinoma (HNSCC) comprises the great majority of HNC cases. With more than half a million cases diagnosed every year, HNSCC is the sixth most common cancer in the world [47].

4.1.3 HNC Risk Factors

Environmental risk factors, namely tobacco use and alcohol consumption, are the most common causes of HNC, which work synergistically, and are responsible for 70–75% of cases [48]. Recently, infection with the human papilloma virus (HPV) has been shown to promote HNSCC and causes cancer. In some parts of Asia, betel-quid chewing has been shown to play a role in the development of HNC [48]. On the molecular level, HNCs often have p53 mutations and many display chromosomal instability. The average age of patients affected with HNC is 63 years; two thirds of these patients are men [49].

4.1.4 HNC in young patients

HNC is traditionally considered to arise in older members of the population; however, during recent years, an increasing incidence of HNC among young adults has been reported. Reports indicate an increase in tumours affecting the tongue and oropharynx among young adults in India, Europe, the USA and China [45]. The growing incidence of HNC in younger patients could be partially explained by a shift in social behaviours and the role of genetics contributing to the development of these tumours. Some authors have proposed that HNC in young

patients might be more related genetic predisposition or HPV infection, compared to HNC in older patients who have had longer exposure to the major risk factors, mainly tobacco and alcohol consumption[45].

4.1.5 Symptoms of HNC

Symptoms of the HNC vary, depending on the site of origin. The five most frequently reported symptoms are: weight loss, pain, feeding difficulties, dysphagia, and cough. It has also been reported that more than one third of HNC patients develop psychological problems [49].

4.1.6 Treatment Options

Treatment of HNC is challenging because of the diversity of the anatomic sites in the head and neck and the critical normal structures around the tumour site, which can result in impairment of vital functions including; breathing, hearing, swallowing, taste and smell [4]. Treatment options for HNC can be definitive surgical resection, chemotherapy, or a combination of both procedures, and radiation therapy. Progress in molecular targeting of HNC has found that cetuximab, an anti-EGFR monoclonal antibody, can benefit patients when it is combined with chemotherapy or radiation [50]. Only a small number of patients have benefitted from the current targeted treatments due to development of drug resistance causing decreased efficacy with long-term treatment. Despite the availability of these aggressive treatments, the 5-year survival rate of for HNC remains relatively poor at 65%, and many patients still develop recurrent tumours and distant metastasis [48]. Therefore, the identification of new therapeutic targets is necessary.

4.1.7 Genetics of HNC

An understanding of the molecular and genetic abnormalities that play roles in oncogenesis of HNC has increased in the past few years due to the development of new methods that allow researchers to study the genome. The development of microarray technology has enabled classification of HNC into distinct types based on gene expression. More recently, the development of NGS has enabled researchers to sequence the whole genome, or whole exome, of a large number of tumours, leading to identification of novel mutations in tumour suppressor genes and oncogenes which can guide to the development of new therapeutics [48].

The studies performed by Stransky et al.[51] and Agrawal et al. [4] in 2011 were the first to sequence the whole exome of HNSCC using NGS. Stransky et al. analysed the whole-exome

sequence from 74 tumours and their matched normal tissues, and Agrawal et al performed the analysis from 34 tumour–normal pairs. Both studies reported that the number of mutations was higher in tumours from patients who smoked than from those who did not smoke. They also found that HPV-negative tumours had higher mutation rate than did HPV-positive tumours, and the difference was independent of patients' smoking status. Both studies reported that the *TP53* gene had the highest number of mutations in HPV-negative tumours, and frequent mutations in cyclin-dependent kinase inhibitor 2A (*CDKN2A*), phosphatidylinositol-4.5-bisphosphate 3-kinase catalytic subunit alpha (*PIK3CA*), and Notch homolog 1 translocation-associated (*Notch1*).

There is a type of HNC emerging in which very young people who have no apparent exposure to alcohol, smoking, or HPV are developing the disease. We thus focused on this group and we were provided with whole exome sequences of genomic DNA derived from the tumours, normal adjacent tissue, and blood, as well as whole tumour RNA, from five HNC patients (clinical features shown in table 4.1). We analysed them using the novel software (CLC-bio), which allows us to assemble fastq files containing next generation DNA or RNA sequencing data to detect tumour specific somatic variants.

No.	Gender	Age at diagnosis	Localisation	Status	Follow-up time
82	Female	19	1	Dead	< 2 years
98	Male	31	3	Alive	> 3years
111	Female	31	2	Alive	> 2 years
119	Male	67	2	Alive	< 2 years
137	Female	71	2	Alive	< 1 year

Table 4.1. Clinical information of the five head and neck cancer patients.
1 = tongue, 2 = border of tongue, 3 = overgrowth outside mobile tongue

4.2 Results

In our research we have analysed the whole exome DNA sequencing of tumours, normal adjacent tissues, and germline blood from five HNC patients (82, 98, 111, 119, 137) using the CLC Biomedical Genomic Workbench 2.5. Two patients (119 and 137) were old, and the other three are young patients with no smoking or alcohol history, and were negative for HPV infection. Four patients (no RNA for 82) have had their tumours RNA sequenced and compared to the tumour DNA sequence to look for expressed mutations in the tumours. It is appreciated that with such small numbers, the results will not be necessarily clinically robust on predicting risk factors for why young people are acquiring HNC, but the dataset allows us to (i) apply novel software and (ii) determine whether blood (germline) or normal adjacent tissue (not necessarily germline) provides the best source of a "normal" reference genome.

4.2.1 DNAseq analysis

For all patients, the DNA sequence was provided from the tumour, normal adjacent tissue (which is 7–10 cm distant from the tumour), and blood (except in patient 82 the sequence was derived from the tumour and blood only). The total RNA sequences from the tumour samples of four patients (98, 111, 119 and 137) were provided. Clinical samples were lysed using commercially available kits to acquire DNA and RNA. These samples were used by a commercial sequencing facility and processed using an Illumina sequencing platform. For the tumour samples the HiSeq2500 PE100-125 kit was used to sequence the whole exomes from gDNA, paired-end 2x100-125 or PE100-125 (read length). The estimated average on-target coverage is 100x. For the normal adjacent tissues and blood, the HiSeq2000/2500 PE100 kit was used, paired-end 2x100 or PE100 (read length). The estimated average on-target coverage is 30x. The HiSeq2500 system uses sequencing by synthesis (SBS) technology. The SBS technology supports massively parallel sequencing using a proprietary fluorescently labelled reversible terminator method that enables detection of single bases as they are incorporated into growing DNA strands. A fluorescently-labelled terminator is imaged as each dNTP is added and then cleaved to allow incorporation of the next base. Since all four reversible terminator-bound dNTPs are present during each sequencing cycle, natural competition minimizes incorporation bias. SBS technology supports both single read and paired-end libraries.

The DNA and RNA sequences of these patients were imported as fastq files and analysed in the CLC biomedical workbench 2.5. Below I will review the general workflow of this application, since it is the first robust software tool that does not require coding to generate outputs. The tool 'identify somatic variants from tumour normal pair (whole exome sequence) ready to use workflow' was used to identify potential somatic variants in tumour samples by using the blood as a control sample. The analysis steps in the software options were presented as follows, in order to describe the options in detail:

STEP1; The **Identify Somatic Variants from Tumour Normal Pair (WES)** ready-to-use workflow is used to identify potential somatic variants in the tumour sample when we have a normal/control sample from the same patient (fig 4.1).

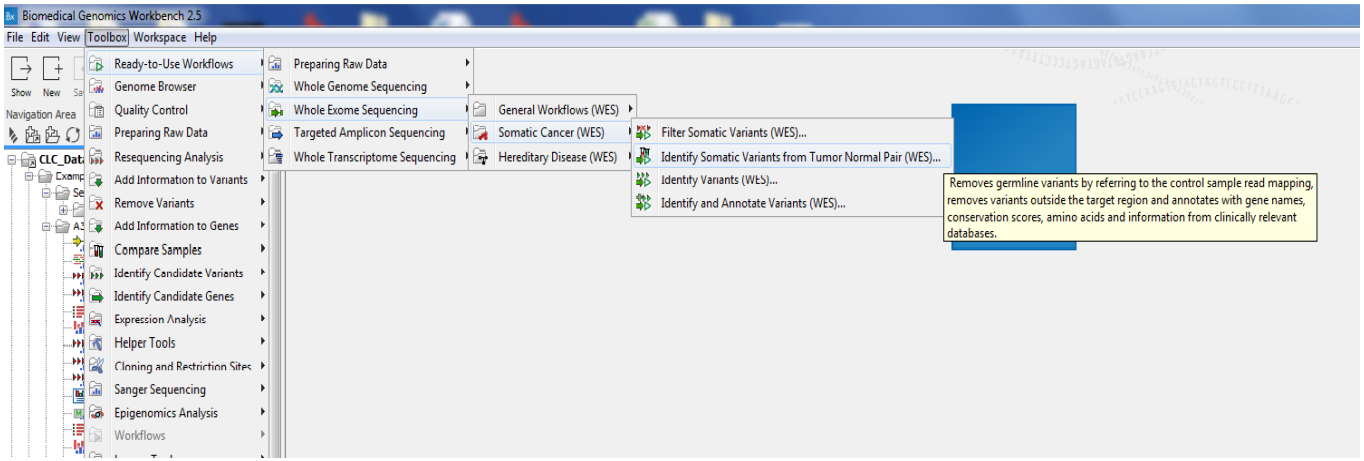


Figure 4.1. Identification of somatic mutations from tumour-normal pair*. From the toolbox options; the ready to use workflow was chosen; then from the five options; whole exome sequencing was chosen; then somatic cancer; then identify somatic variants from tumour normal pair. Identify Somatic Variants from Tumour Normal Pair: Removes germline variants by referring to the control sample read mapping, removes variants outside the target region (outside exomes), and annotates the output files with gene names, conservation scores, amino acid changes, and information from clinically relevant databases. WES: whole exome sequence.*

STEP 2; The paired-end tumour sequencing reads (previously imported from fastq files, named below as “index26” were selected from the navigation area as shown in figure 4.2.

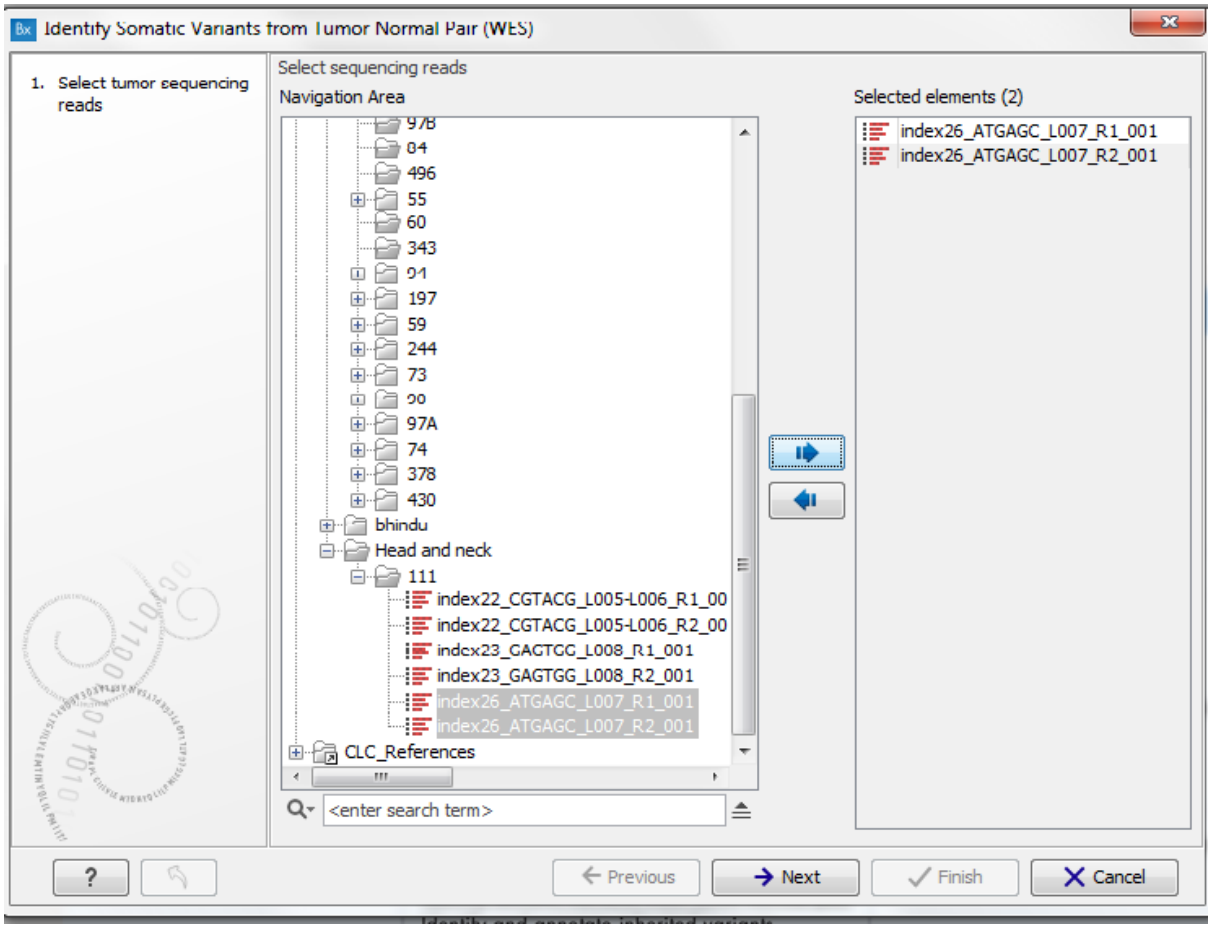


***Figure 4.2**. **The tumour reads from each paired-end read were selected**. Samples are in patient 111 file (as in folder in the left column). The index26 represents the tumour WES. On the right side of the table, it highlights the selected DNAseq files from the paired-end sequencing (R1 and R2). This information is then transferred to the next step of the work flow: Figure 4.3.*

STEP 3; the paired-end normal sequencing reads (previously imported from fastq files, named below as “index22…”); the normal sequencing reads were selected as in figure 4.3.

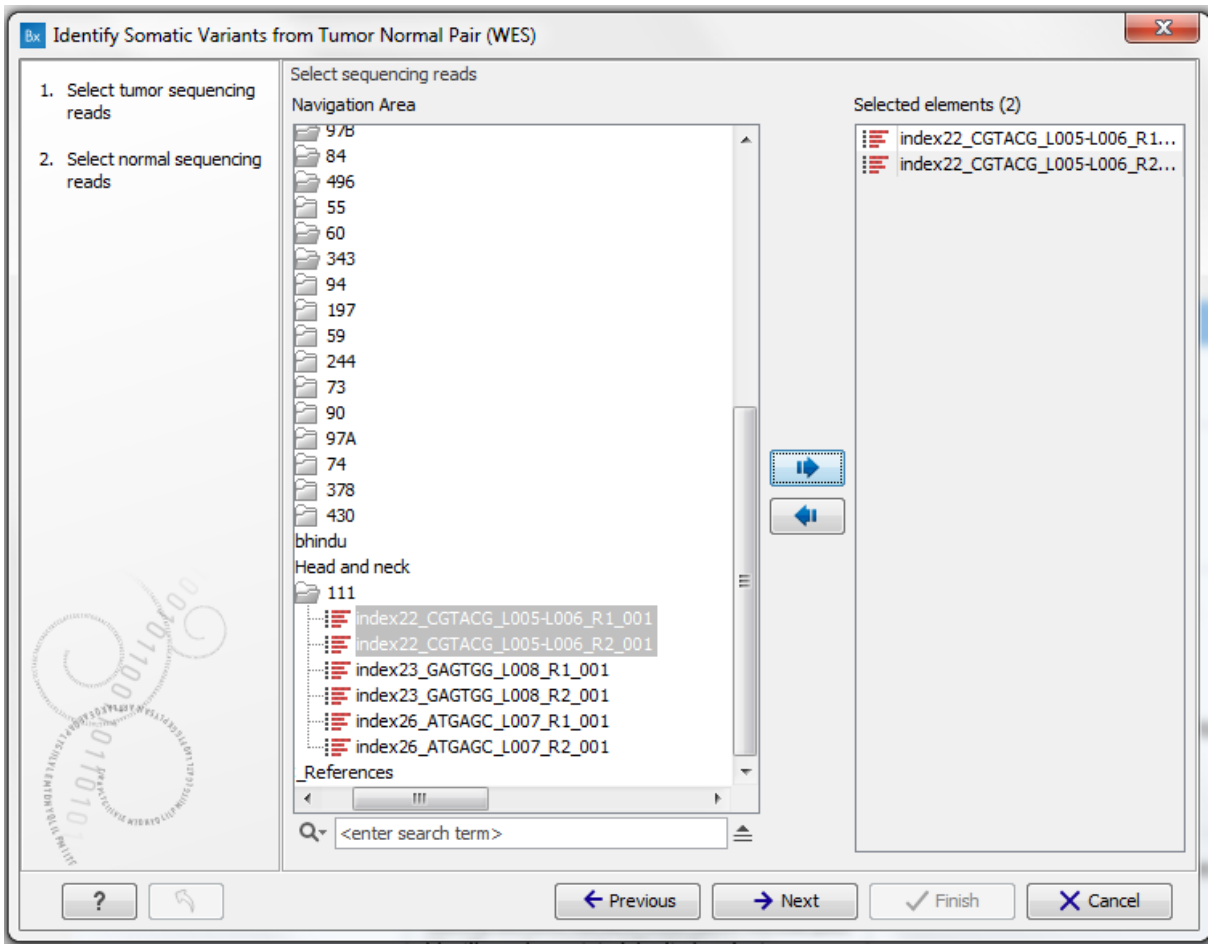


***Figure 4.3**. **The normal reads were selected**. Samples are in patient 111 file (as in folder in the left column). The index22 represents the normal WES. On the right side of the table, it highlights the selected DNAseq files from the paired-end sequencing (R1 and R2). This information is then transferred to the next step of the work flow: Figure 4.4.*

STEP 4; The targeted region file was added (figure 4.4), which is the file that specifies which regions have been sequenced using the Illumina exome sequencing primers, when working with whole exome sequencing data.

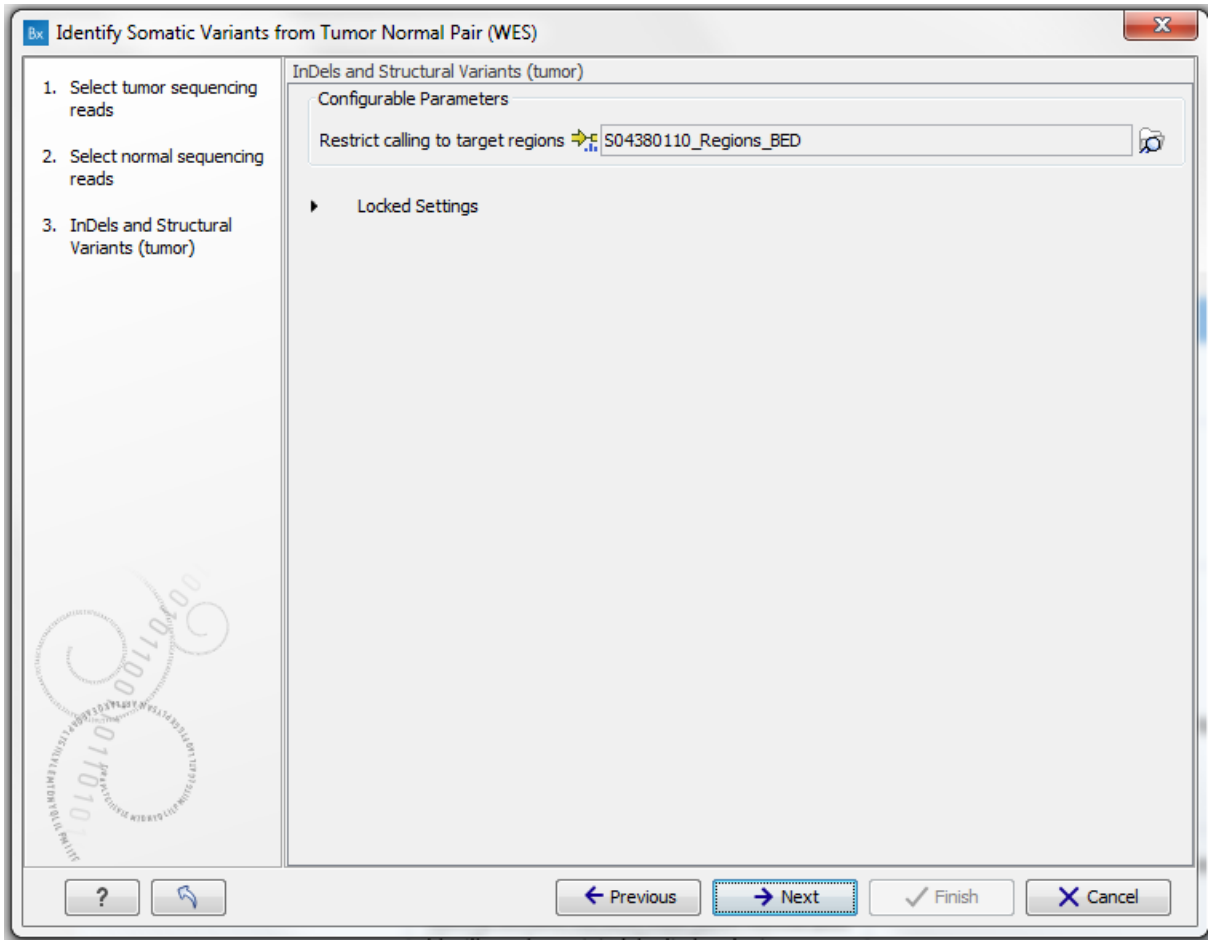


Figure 4.4. Addition of a target region*. A file with the genomic regions targeted by hybridization kit is available from the vendor of the enrichment kit and sequencing machine. The file is in bed format.*

STEP 5; The parameters for minimum coverage, minimum count and minimum frequency for the detected mutation, could be changed as shown in figure 4.5. These are the parameters that are usually altered if computer coding is used in standard open source software. The CLC-bio software provides an interface that allows manual alterations in these parameters without knowledge of computer code. We had chosen the minimum coverage to be 5, so only variants in this region covered by at least 5 reads were called. We set the minimum count to 2, so only variants that were present in at least 2 reads were called. The minimum frequency (count/coverage) was set to 5, so only variants that were present in at least 5% frequency were called.

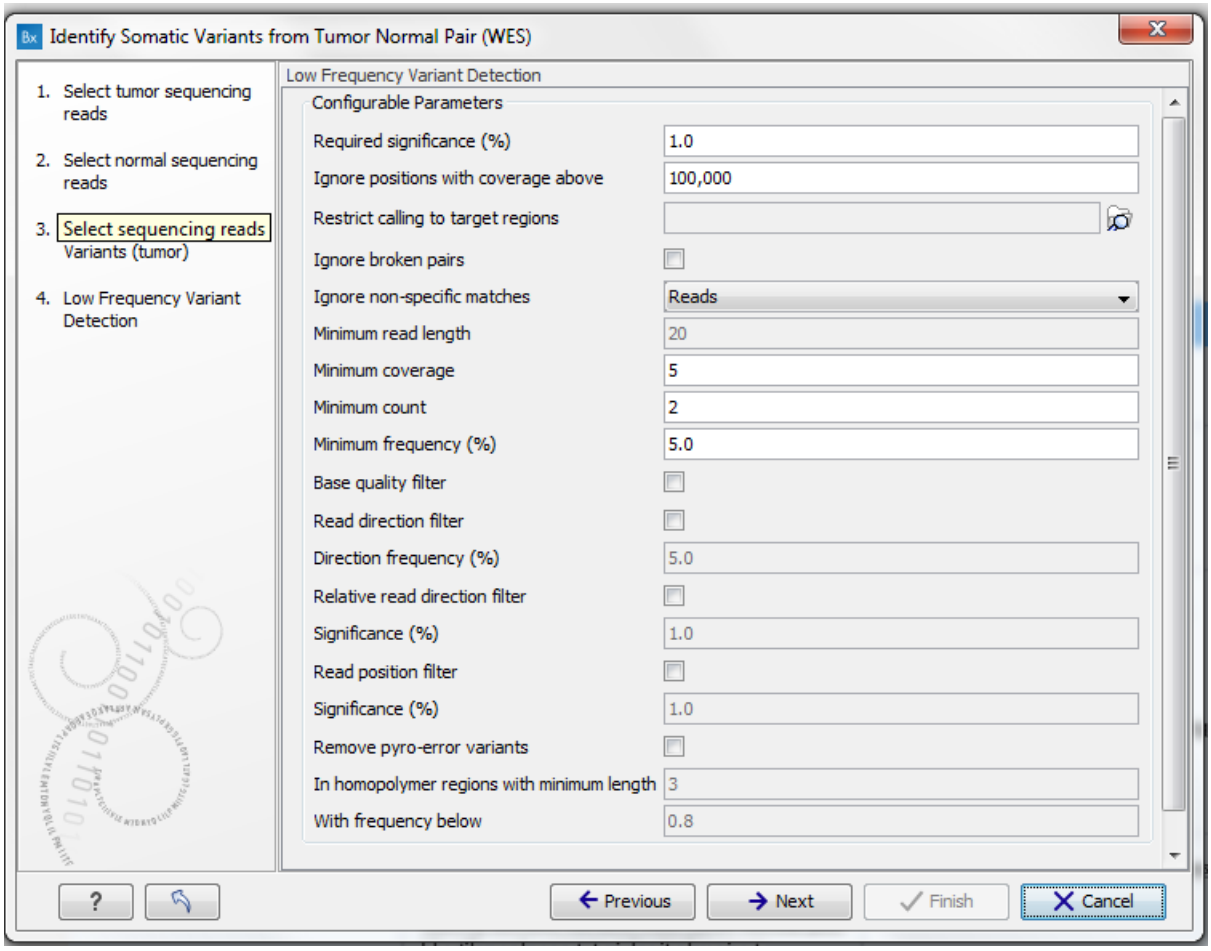


Figure 4.5. Set the parameters for the analysis*. The parameters for the variant detected are adjustable. The minimum coverage of any variant detected will be 5 reads in the tumour, and the minimum counts are 2, and the minimum frequency is 5%. Any variants having lower of any of these parameters would not be called out.*

STEP 6 – The aim of this step is to define only tumour-specific variants (mutations) only, so the parameter selected is 'keep variants with control reads below' 1, (figure 4.6). This means any variant that presents in at least one read in the control would not be called.

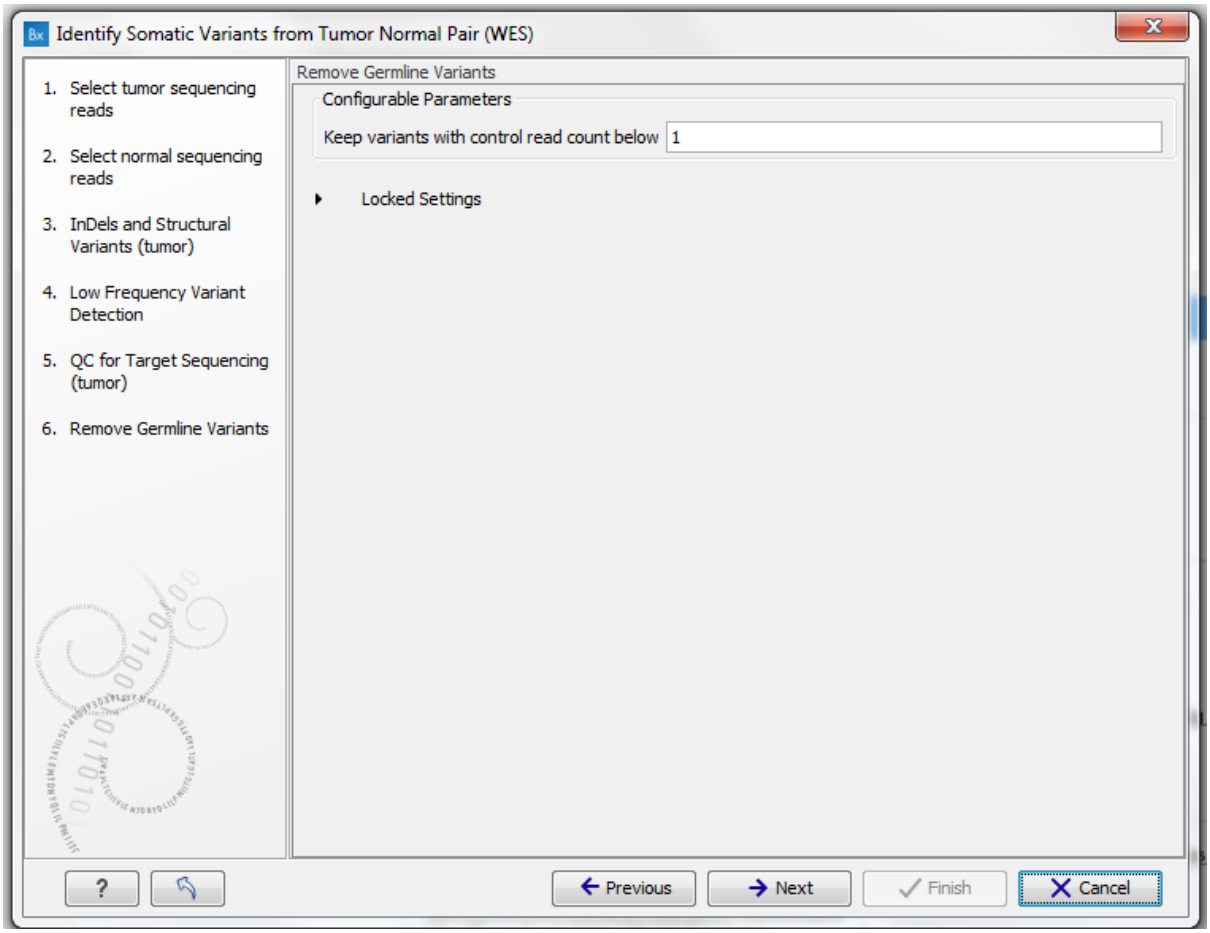


Figure 4.6. Adjust the settings for removal of germline variants step*. Keep variants with control read count below 1, so any variant that presents in at least one read in the control would not be called.*

STEP 7- In the next step of the work-flow, the same target region for the normal sequencing is chosen as in the tumour; to make sure there are enough sequencing reads in the normal sample at the mutation site, we set the minimum coverage to be 5 in the normal reads (figure 4.7).

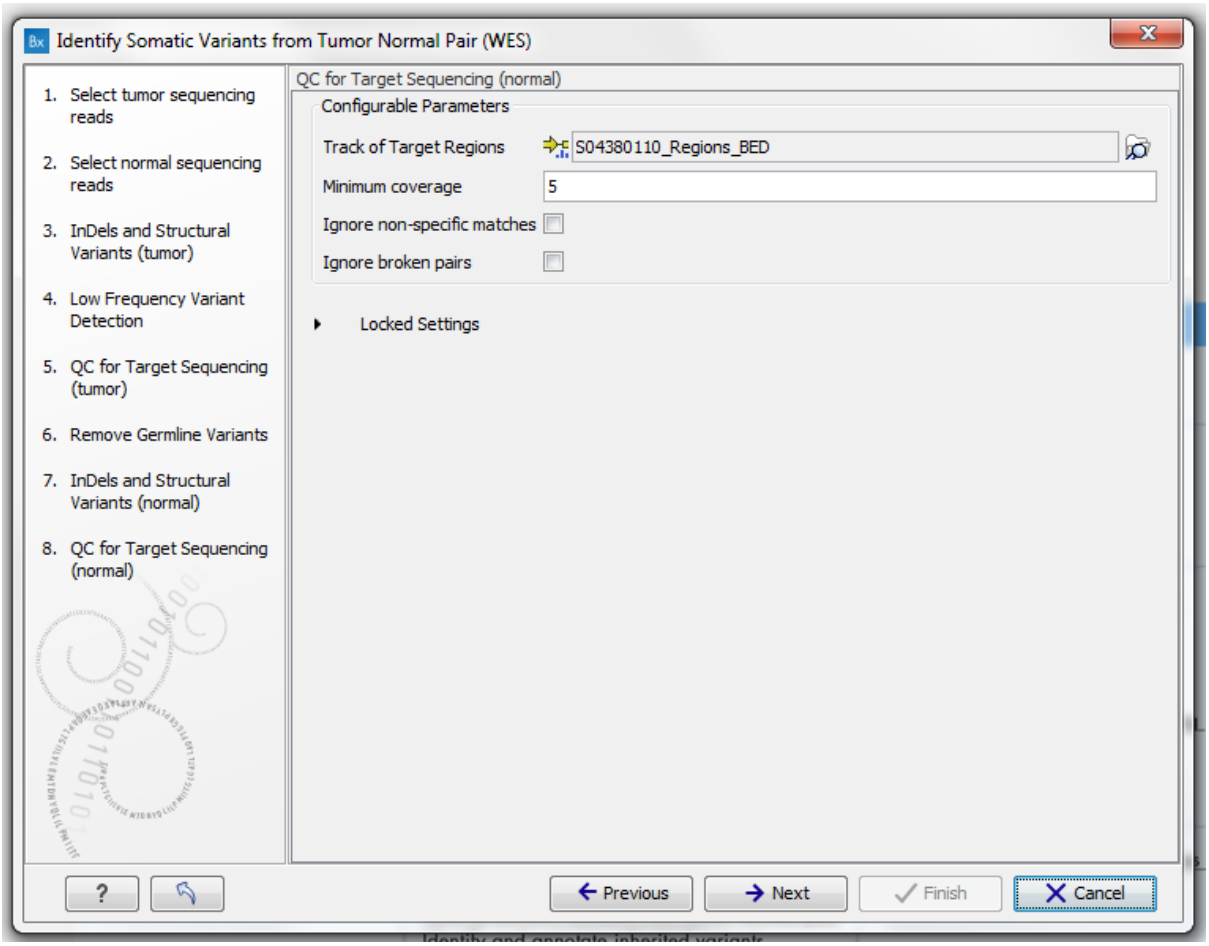


Figure 4.7. Select the target region for the normal sequencing and set the minimum coverage in the normal reads.

STEP 8- Save the file (fig 4.8) and the computational analysis has initiated.

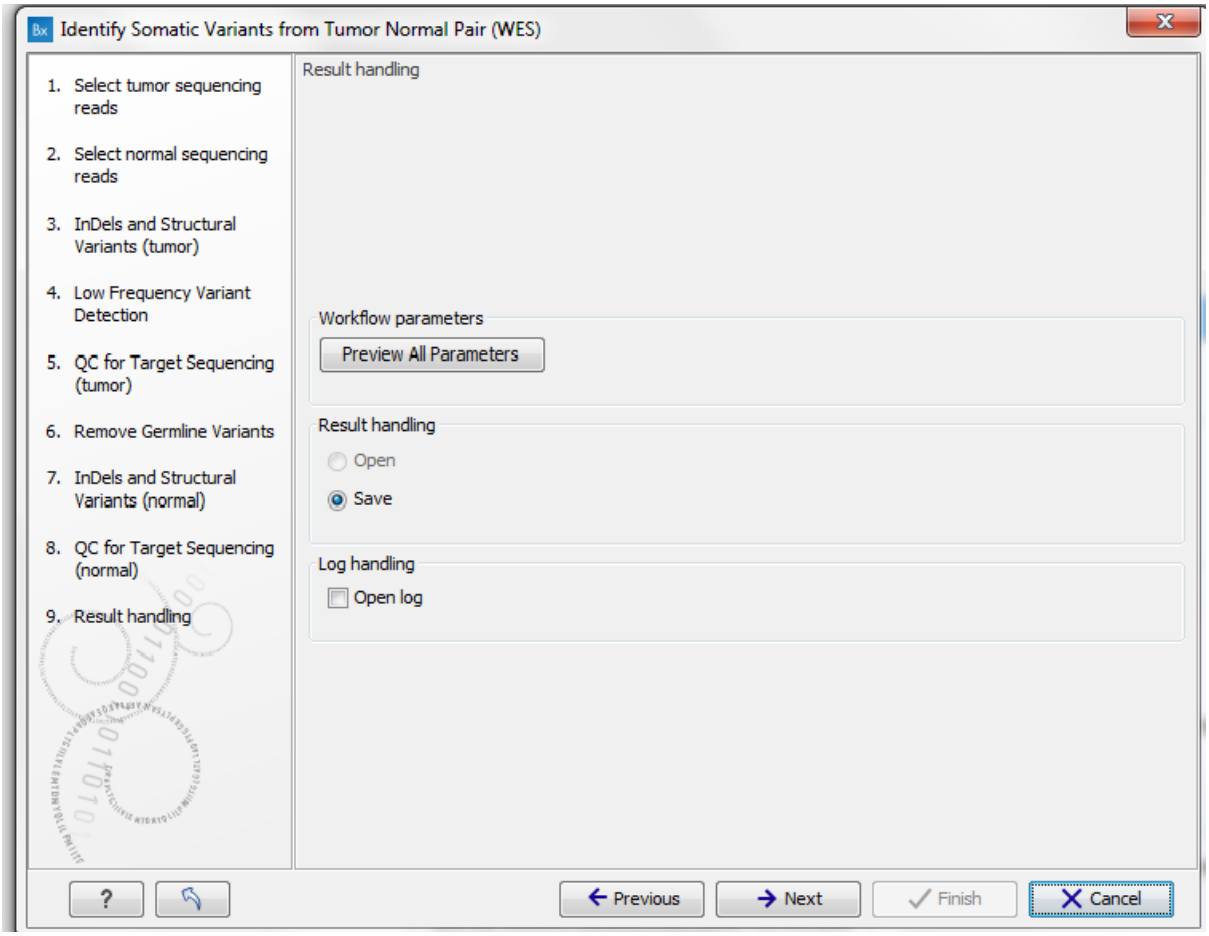


Figure 4.8. Check the parameters and save the results.

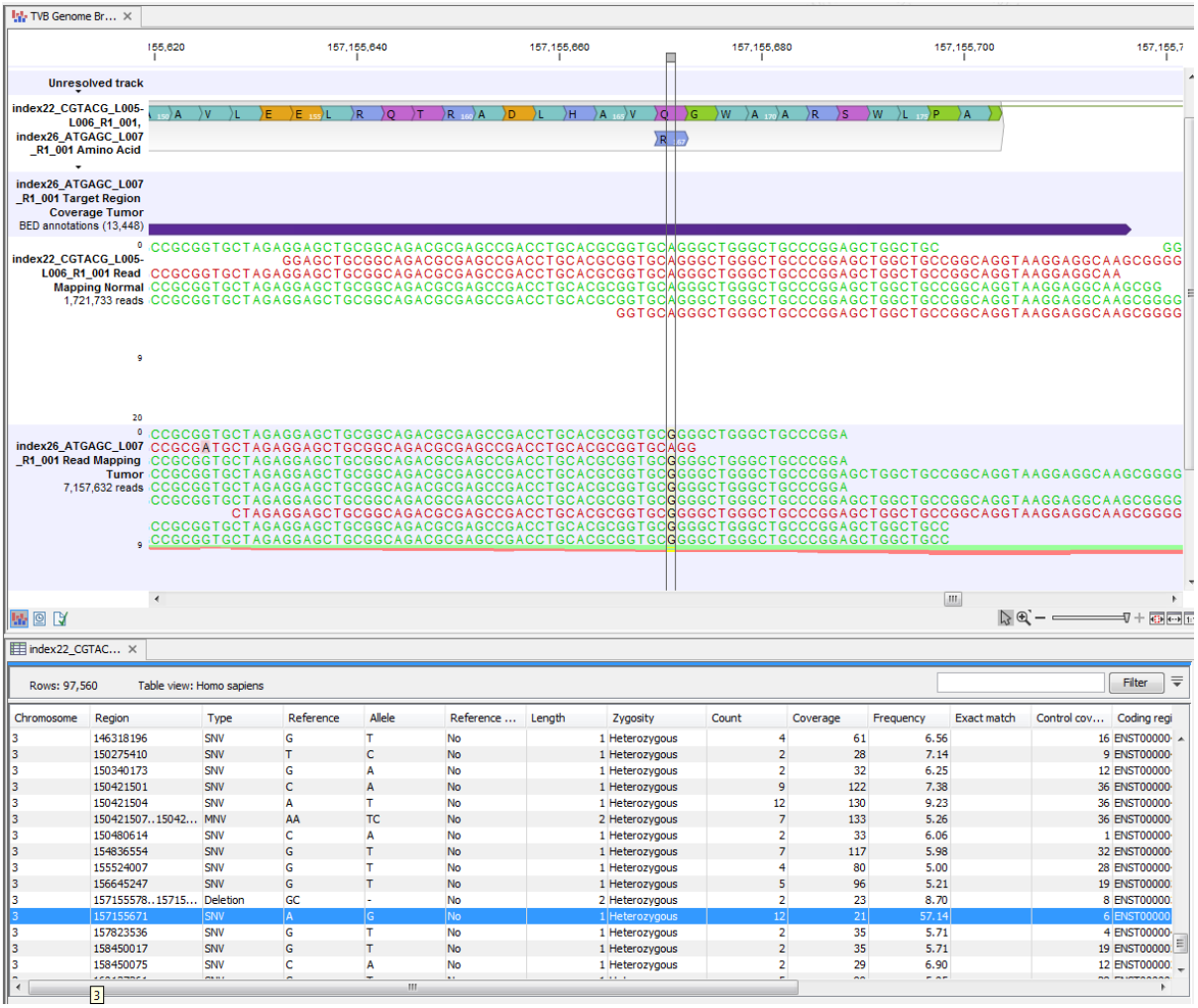
4.2.2 Analysis summary

When the workflow was run, the DNA sequencing reads were mapped and the variants (mutants) identified. An internal workflow removed germline variants that were found in the mapped reads of the control sample, and variants outside the target region were removed as they were likely to be false positives due to non-specific mapping of sequencing reads. The remaining variants were then annotated with gene names, amino acid changes, and information from clinically relevant databases like ClinVar (variants with clinically relevant association).

The analysis results in an output table with all the detected variants and their details, and genome browser as shown in figure 4.9. If any variant were selected from the table, it would appear on the browser, which shows the tumour reads with the mutation, and the normal reads, and the amino acid change. The sequencing reads are shown in different colours

depending on their orientation, whether they are single reads or paired reads, as explained below:

- The colour of the consensus and reference sequence. Black per default.
- Forward: The colour of forward reads (single reads). Green per default.
- Reverse: The colour of reverse reads (single reads). Red per default.
- Paired: The colour of paired reads. Blue per default, in this case, because the original fastq files were not imported together as paired-end reads, then only individual strands are analysed and there is no fusion to produce paired-end reads (blue). The choice to not use pair-end read import allows visual analysis of each read (forward or reverse) for examination.



Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Coverage	Frequency	Exact match	Control cov...	Coding regi
3	146318196	SNV	G	T	No	1	Heterozygous	4	61	6.56		16	ENST00000
3	150275410	SNV	T	C	No	1	Heterozygous	2	28	7.14		9	ENST00000
3	150340173	SNV	G	A	No	1	Heterozygous	2	32	6.25		12	ENST00000
3	150421501	SNV	C	A	No	1	Heterozygous	9	122	7.38		36	ENST00000
3	150421504	SNV	A	T	No	1	Heterozygous	12	130	9.23		36	ENST00000
3	150421507..15042...	MNV	AA	TC	No	2	Heterozygous	7	133	5.26		36	ENST00000
3	150480614	SNV	C	A	No	1	Heterozygous	2	33	6.06		1	ENST00000
3	154836554	SNV	G	T	No	1	Heterozygous	7	117	5.98		32	ENST00000
3	155524007	SNV	G	T	No	1	Heterozygous	4	80	5.00		28	ENST00000
3	156645247	SNV	G	T	No	1	Heterozygous	5	96	5.21		19	ENST00000
3	157155578..15715...	Deletion	GC	-	No	2	Heterozygous	2	23	8.70		8	ENST00000
3	157155671	SNV	A	G	No	1	Heterozygous	12	21	57.14		6	ENST00000
3	157823536	SNV	G	T	No	1	Heterozygous	2	35	5.71		4	ENST00000
3	158450017	SNV	G	T	No	1	Heterozygous	2	35	5.71		19	ENST00000
3	158450075	SNV	C	A	No	1	Heterozygous	2	29	6.90		12	ENST00000

Figure 4.9. The results of the analysis*. The table shows all the variants detected with many details of each variant such as the chromosome number, the region and type of the variant, the count, coverage, frequency, control coverage, gene name and amino acid change. The file above the table shows the normal and the tumour reads. When any variant is selected, it will appear on the reads above. The selected variant shows that there is a mutation from A to G, which changed the amino acid from Q to R.*

4.2.3 Defining the number of non-synonymous mutations and types of mutations by comparing tumour to blood (germline)

We were interested primarily in defining in non-synonymous mutations only the variants detected towards identifying functional mutated protein drivers. Table 4.2 shows the number of the somatic non-synonymous mutations detected in each patient, and the type of these mutations. Young patients have less than half the number of mutations detected in older patients, except for patient 98, who had a high number of mutations, similar to that seen in the older patients. These mutations were in a very large number of genes. For each variant detected, the table produced by analysis gives all the information, including the count, the coverage, frequency, chromosome number, region, gene name, and amino acid change. By using the chromosome number and the variant region, the variant could be viewed in the browser as shown in figure 4.10, highlighting some mutations in patient 82.

Sample	No. of Non-synonymous mutations	No. of SNV	No. of MNV	No. of Indel	No. of replacement
82	3460	3110	57	288	5
98	7467	6829	71	559	8
111	3302	2943	71	286	2
119	10853	9765	107	975	6
137	8273	7522	63	682	6

***Table 4.2**. **Number of non-synonymous mutations in each patient, and their types**. SNV: single nucleotide variant. MNV: multiple nucleotides variant. Indel: insertion and deletion. Replacement when a number of bases are replaced by more or less number of bases*.

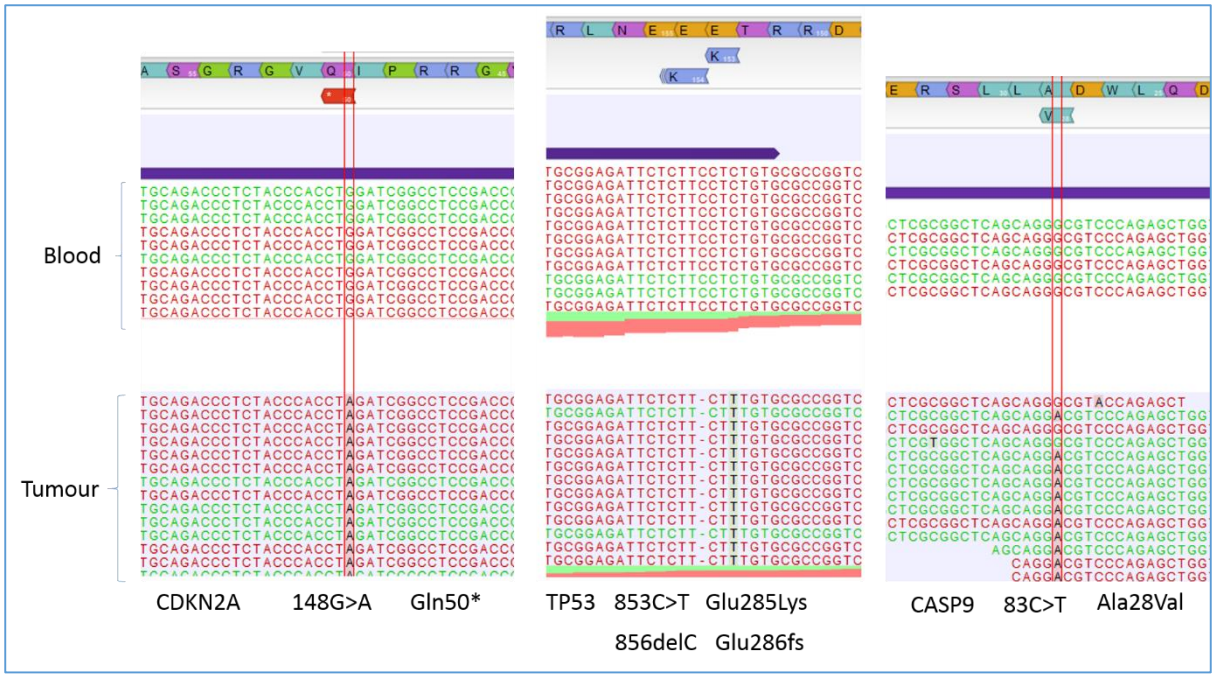


Figure 4.10*. *A genome browser view of the sequencing reads in the indicated genes. *CDKN2A, TP53, and CASP9 mutations in patient 82. The above reads are the normal reads (from blood) and the lower reads are the tumour reads. The amino acid sequence of the protein is shown above, with the amino acid changed at the mutation site or with * where there is a truncation mutation as in the mutation in CDKN2A.*

4.2.4 Defining the mutation signature by comparing tumour to blood (germline)

Somatic mutations are a major cause of cancer development. Each of these mutations was caused by the activity of endogenous and/or exogenous mutational processes with different strengths. Understanding the mutational mechanisms in cancer can yield further insights both into the biology of cancer cells, and of cellular processes in general regulating DNA damage and repair.

In lung cancer, tobacco smokers have on average a 10-fold increase in the burden of somatic mutations in their genomes compared to non-smokers. This is mainly due to the increase in the number of C>A transversions, which is consistent with the experimental evidence for tobacco carcinogens [27]. In glioblastoma, it was demonstrated that treatment with an alkylating agent, such as temozolomide, significantly elevates the number of somatic mutations and results in a distinct mutational pattern of C>T transitions [27]. We wished to define mutation spectra in our head and neck samples to see what type of mutation signature they have, and whether there are any differences between the old and young patients. When we looked at the types of mutations of the SNVs, all five patients have the same type of mutations, with C>A or G>T being the highest number, then C>T or G>A, as shown in figure 4.11 below.

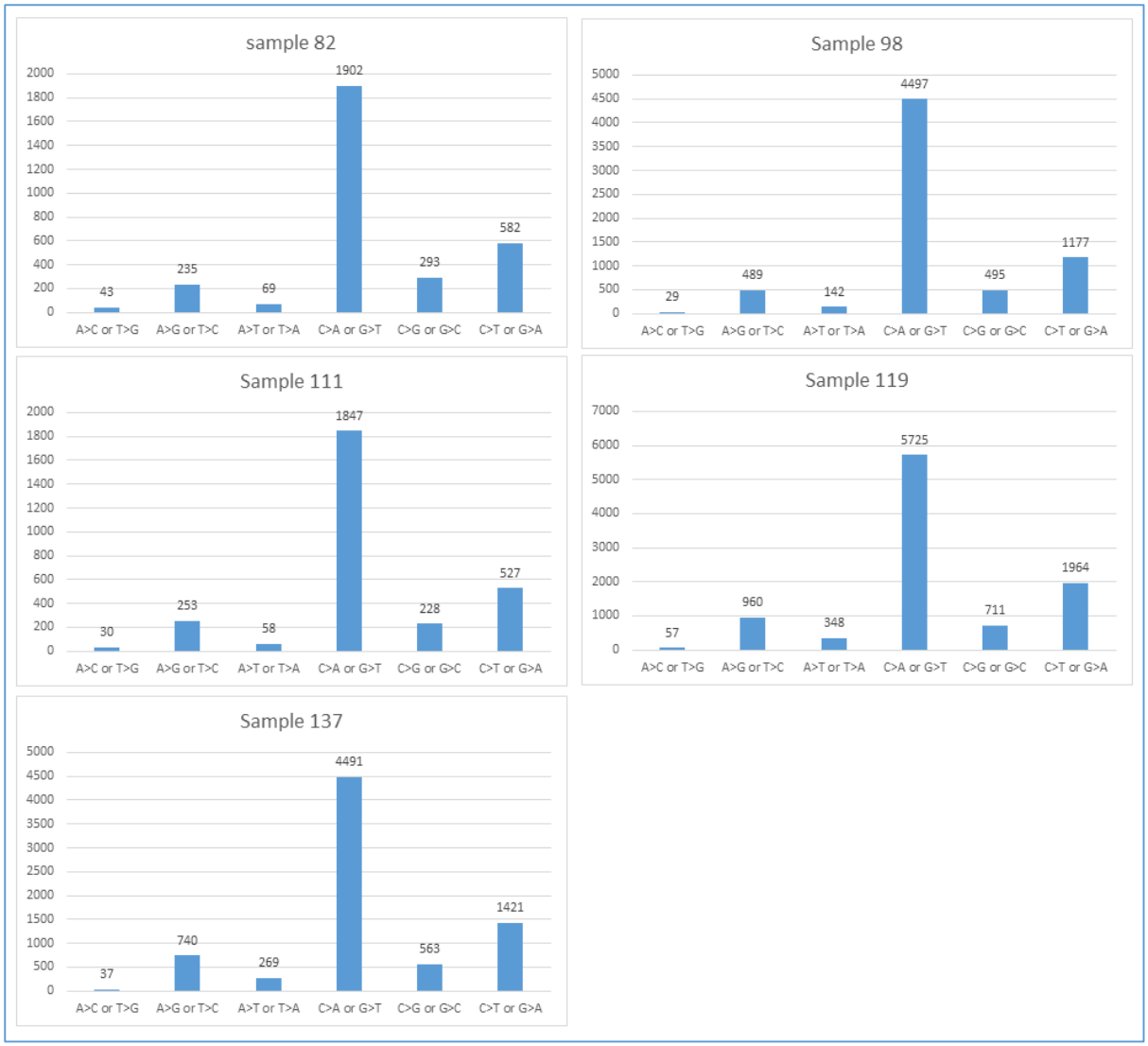


***Figure 4.11**. **The number of each mutation type of single nucleotide variants (SNVs) in all the five patients**. All patients have C>A as the commonest mutation type.*

4.2.5 Defining the commonly mutated genes

The number of genes detected in each patient is very high. In order to lower this number; we have applied some cut-offs to increase stringency. Genes that were detected in at least two patients and had mutations with ≥ 15% frequency (count/coverage) have been chosen for discussion (table 4.3).

Number		Young patients			Old patients	
1	Gene name	82	98	111	119	137
2	RARRES1, MFSD1, RP11-379F4.4	√	√			
3	SRRD	√	√	√		√
4	SHROOM4	√	√	√	√	
5	TNRC18	√	√	√		√
6	ANKLE1	√	√	√		√
7	EP400	√	√		√	√
8	MUC5B	√	√	√	√	
9	MMP23B	√	√		√	
10	MUC4	√	√	√	√	√
11	NBPF1	√	√	√		
12	PKD1	√	√			
13	MACROD1	√	√			
14	ENTPD1	√	√			√
15	RP11-1055B8.7	√	√			
16	LOR	√	√			
17	GOLGA6L2	√	√	√	√	√
18	KRTAP4-7	√		√		√
19	RP11-458D21.5, NBPF10	√		√	√	
20	MESP2	√		√		√
21	FXN	√		√	√	
22	NBPF20	√		√		
23	KRTAP4-11	√		√	√	√
24	MAML2		√	√		√
25	TPRXL		√	√		
26	CRIPAK		√	√	√	√
27	KRTAP10-10		√	√	√	
28	KIAA0430		√	√		
29	ADAM12		√	√		
30	POLRMT		√	√		
31	PCDHB2		√	√		
32	TPRX2P		√	√		√
33	GSPT1				√	√
34	DHFR, MSH3				√	√
35	SELO				√	√
36	ENAH				√	√
37	GOLGA8B		√		√	√
38	UGT2A1				√	√
39	FAM194B	√			√	√
40	DSPP		√		√	√
41	LRP2				√	√
42	HERC2				√	√
43	MUC12	√			√	√
44	BRWD1				√	√
45	MCC				√	√
46	KRTAP4-11	√			√	√
47	RP11-683L23.1	√			√	√
48	SPINT2, CTB-102L5.4				√	√

49	TP53	√			√	√
50	RAB36	√			√	
51	KRT10	√			√	
52	DACT3	√			√	
53	COMTD1	√			√	
54	CD163L1	√			√	
55	GAREML	√			√	
56	KRTAP4-5	√			√	
57	FOXC2	√			√	
58	KRTAP10-4	√			√	
59	EVI5L	√			√	
60	COMMD1	√			√	
61	C2CD4A	√			√	
62	ZFPM1	√			√	
63	CASP9	√				√
64	KRTAP4-8	√				√
65	CDKN2A	√				√
66	TNRC6A	√				√
67	C10orf71	√				√
68	KIF11	√				√
69	WDR7	√				√
70	MICALL1	√				√
71	ATXN2	√				√
72	TUBB8P7	√				√
73	HCN2	√				√
74	KCNN3		√		√	
75	RP11-108K14.8, PAOX		√		√	
76	KBTBD11		√		√	
77	KRTAP9-4		√		√	
78	BARX1		√		√	
79	GLTSCR1		√		√	
80	KRTAP5-AS1, KRTAP5-2		√		√	
81	VPS37D		√		√	
82	C9orf172		√		√	
83	TRIO		√		√	
84	GAS1		√		√	
85	RPL14		√		√	
86	PRB4		√		√	
87	RP1L1		√			√
88	KRTAP10-2		√			√
89	FAM46A		√			√
90	FLG		√			√
91	CNTNAP3		√			√
92	SCAF1		√			√
93	PPP2R3B		√			√
94	EMILIN3		√			√
95	SSUH2		√			√
96	CACNA1I		√			√
97	GP1BA		√			√
98	KRT1			√	√	
99	AGAP5, RP11-464F9.1			√	√	
100	CTBP2			√	√	
101	C20orf194			√	√	
102	GPLD1, ALDH5A1			√	√	
103	AC090616.2			√	√	
104	TPSB2			√	√	
105	CDC42EP1			√		√
106	NUTM2F			√		√
107	CROCC			√		√

108	SPANXD			√		√
109	KIAA0040			√		√
110	TMEM52			√		√

Table 4.3. A list of genes detected with somatic mutations in at least two patients and ⩾15% mutation frequency (count / coverage).

4.2.6 Identifying expressed mutations by comparing the tumour RNA to the tumour DNA

One problem with using genetic maps to identify drug targets is that there is no guarantee that the mutant gene is expressed at the time of disease presentation. DNA sequencing can identify thousands of genomic variants in a single cancer genome, and even more if the sample is highly heterogeneous. Therefore distinguishing a genomic variant that causes a selective growth or survival advantage to the tumour is a challenging task [52]. The detected somatic variant frequently occurs on a single allele. The impact of a heterozygous mutation will depend on whether the mutation-containing allele is transcribed to RNA. The mutation detected in the genomic DNA may not be transcribed into RNA if the wild-type allele is selectively transcribed. Furthermore, the mutation-containing transcript could activate RNA surveillance mechanisms and cause rapid degradation of the mutation-containing transcript. Nonsense-mediated decay (NMD) surveillance, for example, scans transcripts for the presence of a premature termination codons before the last exon and, when found, initiates degradation of such transcripts [53]. Therefore, defining “active” somatic genome mutations using RNAseq is very important for judging if the DNA mutation is expressed and/or advantageous to tumour development, or not, at the time of disease presentation.

To define for the expressed genomic mutations, we have compared the variants between the tumour DNA and the RNA. The tumour RNA sequence was compared to the tumour DNA sequence in each patient (98, 111, 119, and 137) to detect the common mutations in DNA and RNA (expressed mutations). In order to detect all the expressed somatic mutations, we have set the parameters of coverage and count of the mutation in the RNAseq to 1. A large number of mutations were detected. Then, we removed all the germline variants found in normal control (blood), as we only interested in the expressed somatic mutations. The steps in comparing the variants in DNA and RNA and detection of expressed somatic mutations are shown below in figures 4.12 to 4.16.

Step 1; the icon was chosen that runs “Compare variants in DNA and RNA” ready-to-use workflow to start the analysis as shown in figure 4.12.

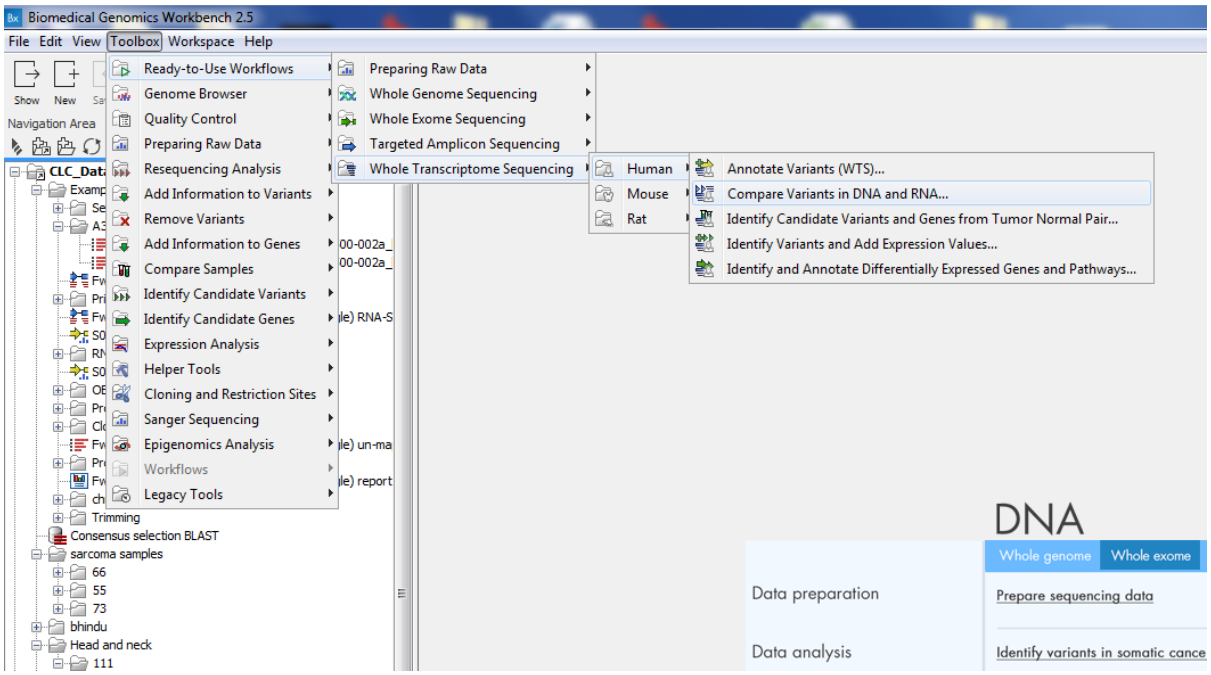


***Figure 4.12**. **Defining variants in DNA and RNA**. The following steps were taken; Toolbox →Ready to use workflows →Whole Transcriptome Sequencing →Human →Compare variants in DNA and RNA. The Compare variants in DNA and RNA ready-to-use workflow identifies variants in DNA and RNA and studies the relationship between the identified genomic and transcriptomic variants.*

STEP 2; After selection of the tumour DNA reads and tumour RNA reads, the parameters were set as shown in figure 4.13. We have set the minimum coverage, count and frequency to 1, because RNA read coverage is not necessarily very high, as shown in figure 4.13 below.

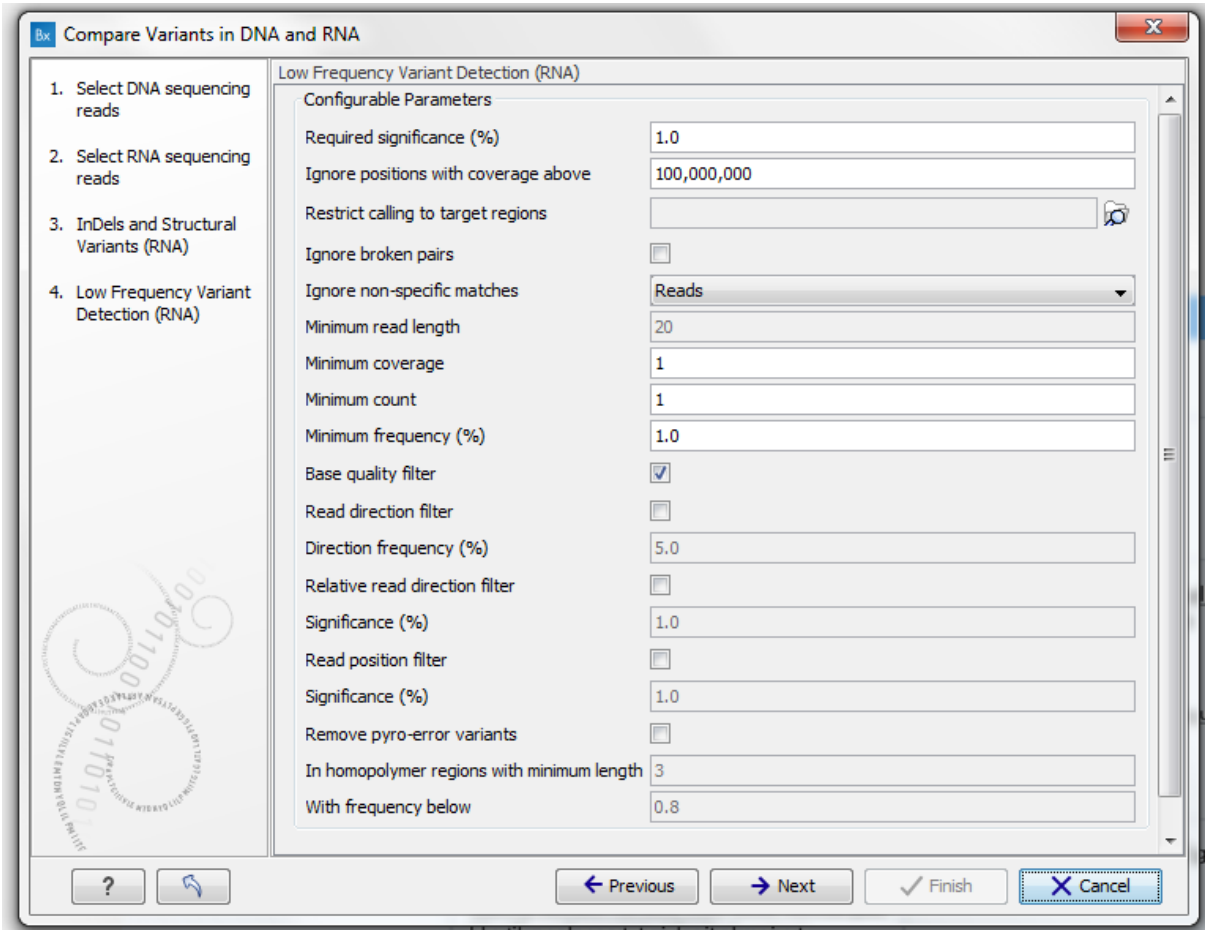


Figure 4.13. ***Set the parameters for the Low Frequency Variant Detection step for the RNA sample***. *Minimum coverage 1, minimum count 1, and minimum frequency 1.*

STEP 3- The samples were filtered against sequencing files that subtract known germline variants in certain populations (fig 4.14).

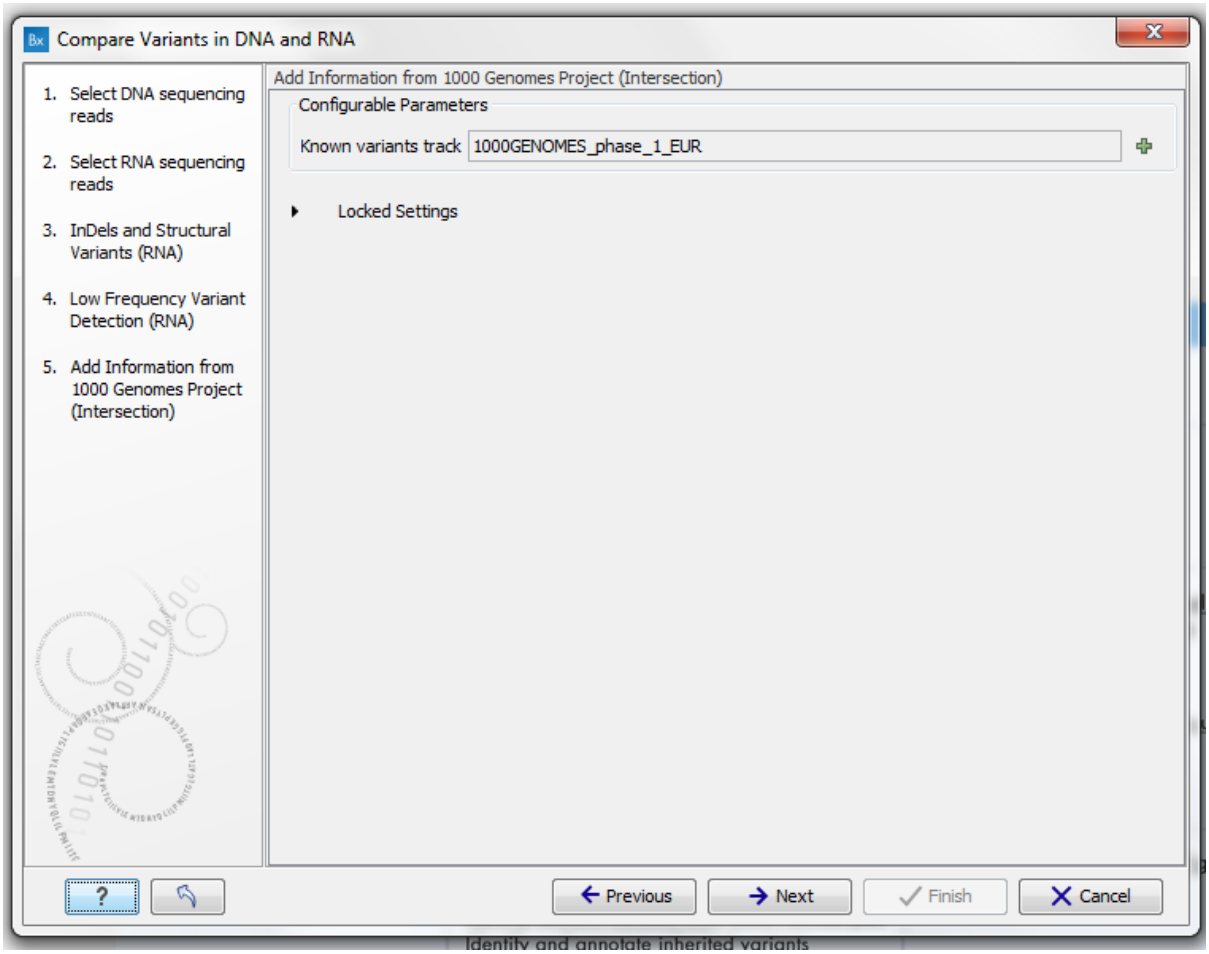


Figure 4.14. ***Specify the relevant 1000 Genomes population for the RNA sample***. *Choose the population that best matches the population the samples are derived from. The European samples have been chosen, as all our samples from Europe*.

STEP 4; The analysis has detected many common mutations in tumour DNA and tumour RNA. Many of these mutations were germline (found in blood). So, in order to get the somatic variants only, we removed all the germline variants that were found in blood as shown in figures 4.15 and 4.16 below:

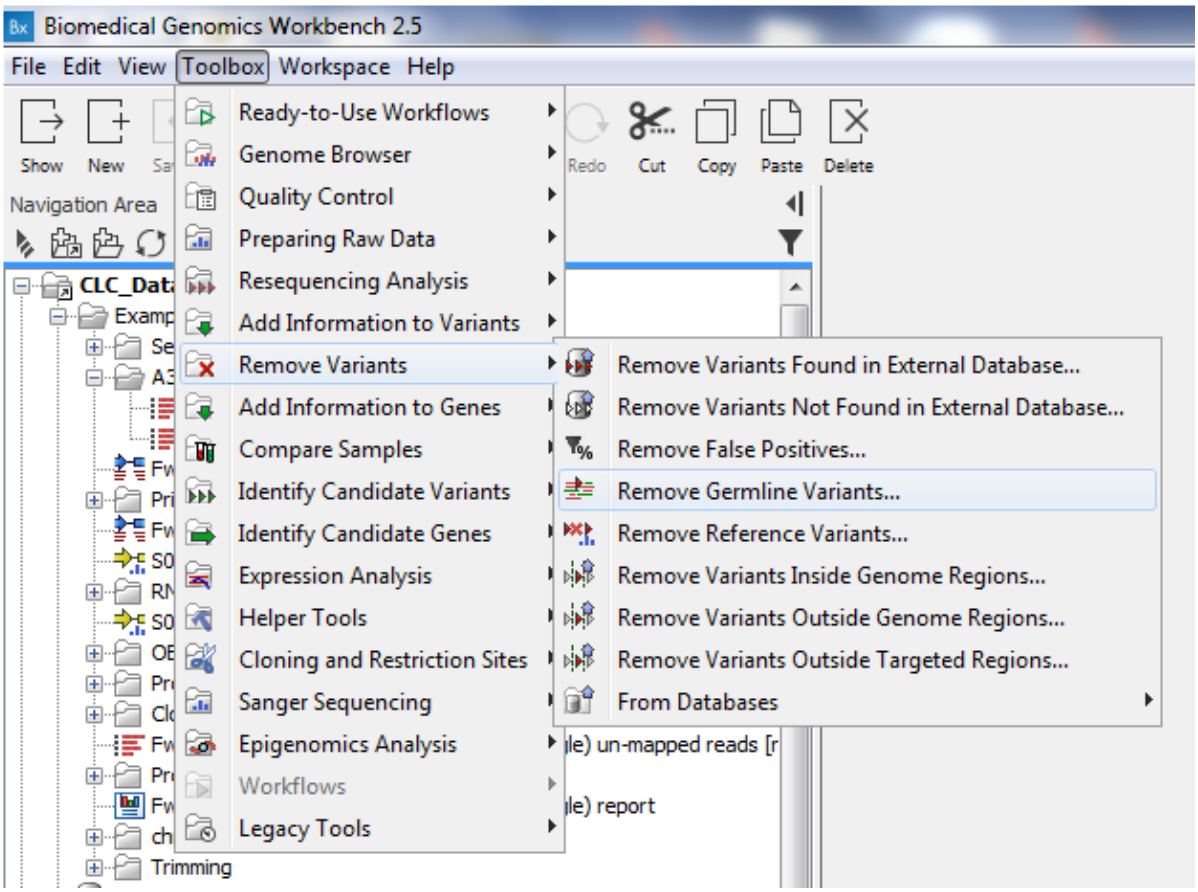


***Figure 4.15**. **Removing germline variants from the expressed mutations which were common in tumour DNA and tumour RNA**. Toolbox → Remove Variants →Remove Germline Variants.*

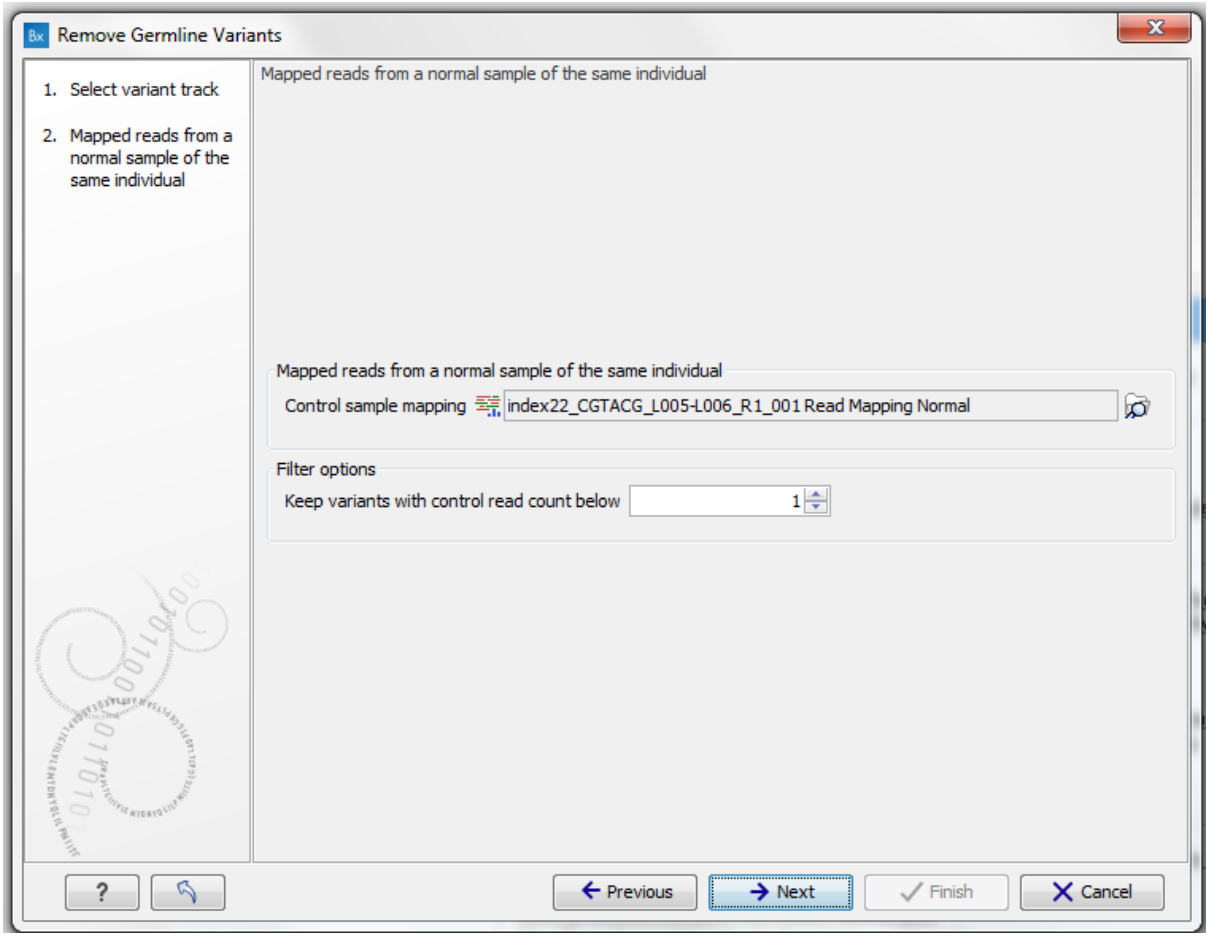


Figure 4.16. Select the normal reads (blood) for each patient to remove the germline variants found in them. *Keep variants with control read count below 1, so any variant in at least one read in germline reads would be removed from our list, leaving us with expressed somatic mutations that only found in tumour DNA and tumour RNA.*

In this workflow, we have removed all the germline variants found in blood, and we are left with expressed somatic mutations specific to the tumour. The number of the expressed somatic mutations in each patient is very low compared to the number of somatic mutations detected in the tumour DNA. This could be due to the low quality of the RNA reads and/or very low coverage compared to the DNA reads as shown in figure 4.17 below, which shows variants with high coverage in tumour DNA but very low coverage in the tumour RNA. Previous studies [54] have shown that only 36% of validated somatic SNVs were observed in the transcriptome sequence when RNAseq data compared with the genomes/exomes data in breast cancer. Similar proportions were also observed in a lymphoma study in which 137 somatic mutations were expressed in RNAseq, out of 329 total somatic mutations [54]. Many variants had been missed because there were no reads in the RNA at the position of the variants. Alternatively, these mutant genes might not be expressed at the time of surgery and resection.

Detected expressed somatic non-synonymous mutations are listed in table 4.4 below with a minimum 5 coverage reads in the DNA reads:

No.	Patient 98		Patient 111		Patient 119		Patient 137	
	Gene name	Mutation	Gene name	Mutation	Gene name	Mutation	Gene name	Mutation
1	*COL4A1*	19G>C	*XAF1*	396C>G	*HADHB*	3_4insACT	*TAF1B*	187delA
2	*USP36*	2874_2879delGAAAAA	*UBXN11*	1114G>A	*IMMT*	883C>A	*IGKV1-17*	97C>A
3	*SRRD*	106G>C	*TRBV10-2*	136_137delTGinsGA	*NOP58*	1562A>G	*TXNDC9*	*293-1370C>A
4	*NBPF1*	3145G>A	*TRBV10-2*	140G>A	*TMEM39A*	*753G>T	*TTC29*	1232T>C
5	*HLA-DQA2*	45_46delTGinsCA	*CTBP2*	20A>G	*HLA-DRB1*	730G>A	*MEF2C*	410C>A
6	*POTEF*	2392C>A	*CTBP2*	22A>T	*HLA-DRB1*	629_630delCAinsTG	*ALDH7A1*	895G>T
7	*HLA-B*	363C>G	*HLA-C*	621A>C	*CCL26*	61G>T	*HLA-DRB1*	317C>A
8	*MAN2B2*	1330A>G	*TRBV10-2*	145A>C	*KMT2E*	1299A>G	*TRAF3IP2*	1613delA
9	*HLA-C*	176G>A	*HLA-DQB1*	755G>A	*DDHD2*	1982A>T	*CALU*	298G>A
10	*HLA-DRB5*	74C>G	*MMS19*	262+651G>A	*TATDN1*	70G>T	*KCTD9*	866C>A
11	*EGFL6*	143C>A	*PTHLH*	524+1383delA	*NRAP*	445G>A	*WHSC1L1*	1626G>T
12	*HLA-DRB5*	84G>C	*ABCF1*	218delA	*SLC2A3*	272G>A	*CDKN2A*	215C>T
13	*PIGT*	1037delC	*TRBV10-2*	155T>A	*LDHB*	830T>A	*CARD17*	7+2028A>G
14	*ASCC2*	541+332G>A	*FRG1*	3_4insAGA	*LRP1*	4219G>A	*CARD17*	7+2014C>A
15	*CEP95*	1225G>T	*MYD88*	473C>A	*OASL*	615G>T	*B4GALNT1*	672_675delTTAC
16	*TBC1D2B*	485C>A	*TMEM43, RP11-434D12.1*	797G>A	*IGHG4*	704A>G	*FGD6*	522A>C
17	*LARP7*	1770delA	*FAM104A*	431_433delGCA	*RPL13*	398delC	*OAS1*	902C>A
18	*TAF1B*	186_187insA	*SLC30A5*	206G>T	*YWHAE*	562C>A	*POSTN*	278delC
19	*ZRANB1*	2126_2127insA	*RPN2*	*84delA	*ZNF257*	266G>A	*PNN*	91G>A
20			*CAMKK2*	*174_*175insA	*LTN1*	952C>T	*IGHV4-61*	104A>G
21			*HLA-C*	299T>A	*IGLV6-57*	398G>T	*UACA*	93G>C
22			*IGHV4OR15-8*	232A>T	*PDZD11*	65C>A	*ZNF720*	*5136_*5137insA
23			*HMGXB4*	*999delA	*FHL1*	226G>T	*TP53*	535C>T
24			*ARHGEF37*	1810A>G	*FGF13*	704G>T	*CCL13*	58C>T
25			*PTK2B*	2983G>A			*SOCS7*	110C>A
26							*GAREM*	1373T>C
27							*KIF16B*	481C>G
28							*APOBEC3A*	443delC

Table 4.4. Expressed somatic non-synonymous mutations in patients 98, 111, 119, and 137, with the gene names*. Coding region change shown above is in the longest transcript of each gene*

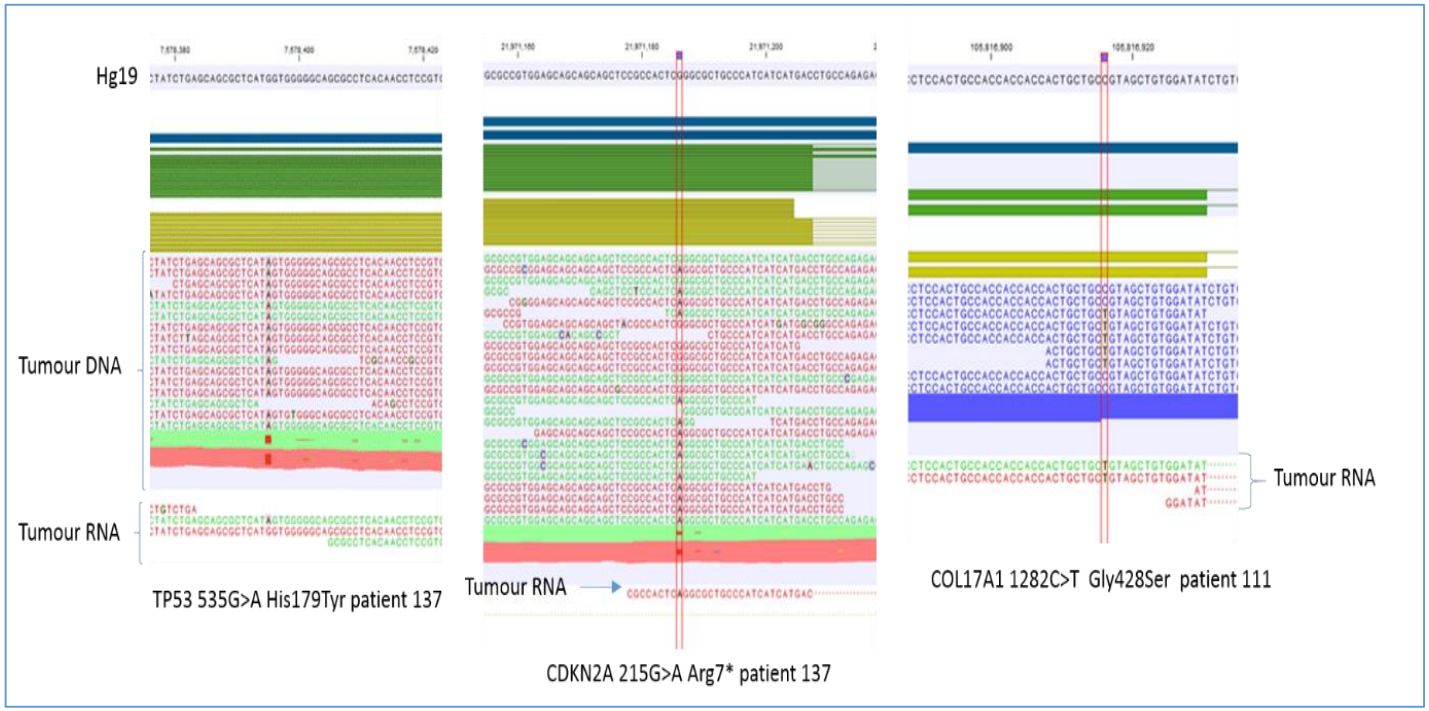


Figure 4.17*. *Examples of expressed somatic mutations in TP53 and CDKN2A in patient 137 and COL17A1 in patient 111. *The variants had high coverage in the tumour DNA reads, but very low coverage in the tumour RNA reads. The reference human genome Hg19 is shown at the top of the image*.

4.2.7 Comparing normal adjacent tissue to blood

4.2.7.1 Number of mutations in normal adjacent tissue compared to tumour tissue

We next compared the DNA mutation landscape in normal tissue versus blood (germline) to determine whether field mutations are dominating. The field cancerization is defined as 'the presence of one or more areas consisting of epithelial cells that have genetic alterations and does not show invasive growth and metastatic behaviour, the hallmark criteria of cancer'[55]. The reads from the normal adjacent tissue have been compared to the blood reads (in the same way as tumour to blood). The number of non-synonymous mutations detected in the normal adjacent tissue is very much higher than that detected in the tumour from each patient as shown in table 4.5 below:

Sample	No. of Non-synonymous mutations in Tumour v Blood	No. of Non-synonymous mutations in Normal adjacent tissue v Blood
98	7467	34558
111	3302	19879
119	10853	24115
137	8273	25652

Table 4.5. The number of non-synonymous mutations detected in the tumour and normal adjacent tissue by comparing them to blood.

4.2.7.2 Mutation signature in the normal adjacent tissue

The mutation type of SNVs in the normal adjacent tissue is similar between the four patients and is exactly the same as in the tumour from each patient, as shown in figure 4.18 below, with the C>A or G>T the highest number.

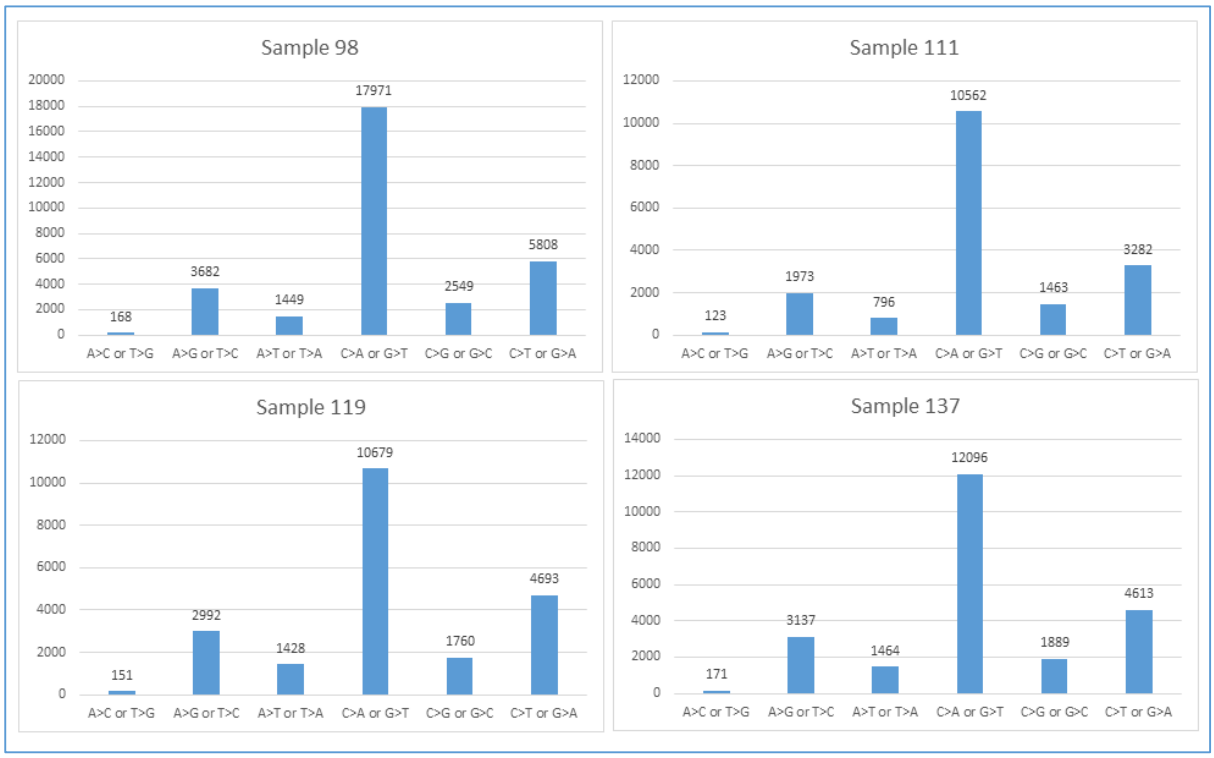


Figure 4.18. The number of mutation types of SNVs in normal adjacent tissue in patients 98, 111, 119, and 137.

4.2.7.3 Common mutations in tumour and normal adjacent tissue

In order to define 'true' cancer mutations, we need to make sure that the mutations detected in the tumour tissue are specific to tumour tissue only. Since we have a high number of field mutations, if we compare tumour vs blood we do not define "true" cancer-specific mutations. As such, we next compared cancer vs "normal" adjacent tissue. This "normal" tissue has a large number of mutations vs germline. This could be due to evolution of a pre-tumour, or it could be that, as we age, "all" of our normal tissue accumulates mutations. As we do not have any other "normal" squamous tissue (from oral mucosa or skin) for comparison we cannot say why the "normal" adjacent tissue has a high number of mutations. However, we believe that a better definition of "true" tumour mutations is a comparison of tumour with "normal" tissue, not "tumour" with blood (germline)". The tumour in each patient shares some mutations with the normal adjacent tissue as shown in figure 4.19 below. These common mutations are listed in the supplementary file. There were some genes in common between the four samples (98,111, 119 and 137) in this list. Table 4.6 below shows genes that were detected in at least two patients, which had same mutation in the tumour and normal adjacent tissue in a patient.

Figure 4.19. Venn diagram showing the numbers of non-synonymous mutations detected in the tumour and the normal adjacent tissue, and the numbers of common mutations between the two tissues.

Number	Gene Name	Sample	Mutation	Amino acid change
1	*CRIPAK*	98	76_77insCA	Cys27fs
		111	76_77insCA 79_80delTGinsCA	Cys27fs Cys27His
		119	205_206delCG	Arg69fs
		137	487G>C	Asp163His
2	*HLA-C*	98	142T>A 97G>C	Ser48Thr Asp33His
		111	621A>C 319G>C 317G>T 313_314delCTinsGC 311A>T	Glu207Asp Gly107Arg Arg106Leu Leu105Ala Asn104Ile
3	*PABPC3*	98	1775T>C 1768T>C 1762C>G	Leu592Pro Tyr590His Leu588Val
		111	1762C>G	Leu588Val
		119	1237_1238delCAinsTG 1228A>T	His413Cys Thr410Ser
4	*KRTAP10-10*	98	83-69751_83-69737delGGGCACACAGCAGAC	Cys105_Cys109del
		119	83-69738_83-69724delACAGGCACACAGCAG	Cys105_Cys109del
5	*GOLGA6L2*	98	1394C>A 1404_1405delGCinsAG	Thr465Lys Gln469Glu
		119	2504G>C 2499_2501delGCC 1576_1577insAGA 2611_2612delGG 1565A>G 2620delA 2624_2625insC 2494_2496delGCAinsATG 2618C>T	Gly835Ala Pro834del Glu525_Met526insLys Gly871fs Asp522Gly Arg874fs Glu875fs Ala832Met Ala873Val
		137	2157_2158insAGATA 2183A>G 1539_1541delATGinsGCA 2629G>A	Ser720fs Glu728Gly Ile513_Trp514delinsMetGln Gly877Arg
6	*FRG1*	98	463T>C 447G>C 495_498delAGAA 482_493delAAGCAGGGGACA 479_480delAT	Cys155Arg Leu149Phe Ile165fs Glu161_Ile165delinsVal Asn160fs
		119	479_480delAT 482_493delAAGCAGGGGACA 495_498delAGAA	Asn160fs Glu161_Ile165delinsVal Ile165fs
7	*TRBV11-1*	98	322A>G	Met108Val

		119	214_216delAGTinsGAA 218T>C	Ser72Glu Val73Ala
		137	4A>G 25A>G	Ser2Gly Met9Val
8	*FAM86C1*	98	385T>C	Cys129Arg
		119	411+49A>G	Thr120Ala
9	*MUC3A*	98	799A>T	Asn267Tyr
		119	56_57delCAinsTG	Ser19Leu
		137	51_52delGTinsAG 43G>T	Trp18Gly Ala15Ser
10	*ANKLE1*	111	*96_*109delGTGTGTGTG TGTGT	Cys586fs
		119	*139_*140delTT	Leu591fs
11	*ZNF66*	111	1056C>G 1518T>G 1514C>G	Tyr352* Asp506Glu Ala505Gly
		119	1045delG	Gly349fs
12	*TRBV10-2*	111	136_137delTGinsGA 140G>A 145A>C	Trp46Glu Ser47Asn Ser49Arg
		119	56G>T 85T>C 46G>A	Arg19Met Tyr29His Ala16Thr
13	*CTBP2*	111	1793C>A	Ala598Asp
		119	1680T>A 1924G>A	Ser560Arg Val642Met
		137	1680T>A	Ser560Arg
14	*MUC16*	111	36513_36515delTGGinsC AA 36520_36522delCAGinsA CT 40847G>A	Gly12172Asn Gln12174Thr Gly13616Glu
		119	40730G>A 40754C>G 37990T>C 38009G>C 37996C>T	Ser13577Asn Thr13585Ser Trp12664Arg Ser12670Thr Pro12666Ser
		137	36780T>G 37017_37019delGAAinsA GG	Ser12260Arg Asn12340Gly
15	*MTCH2*	111	235_236delTGinsCA 229A>T 202C>T	Cys79His Arg77* Arg68Cys
		119	202C>T 176A>T 195_196delTGinsCA 178C>A	Arg68Cys Gln59Leu Gly66Arg His60Asn
16	*MUC5B*	111	9569G>A	Arg3190Gln
		119	6059C>T	Ala2020Val
17	*NBPF1*	111	2037C>A	Asp679Glu
		119	2077G>T	Gly693Trp

18	*CT47B1*	111	319G>T	Asp107Tyr
		119	306G>T	Glu102Asp
19	*KRTAP10-2*	98	303+16587C>T	Pro87Leu
		137	303+16895_303+16909delCCTGTCTGCTGCAAG	Pro190_Lys194del
20	*USP15*	98	454delG	Glu152fs
		137	450G>T	Arg150Ser
21	*SRRD*	111	106G>C	Gly36Arg
		137	106G>C	Gly36Arg
22	*HERC2*	119	821C>G 839G>T 836delG	Ala274Gly Ser280Ile Gly279fs
		137	815_816delCGinsTA 821C>G 836delG 839G>T	Thr272Ile Ala274Gly Gly279fs Ser280Ile
23	*MUC4*	119	11245_11246delAGinsGA 5786C>T 7093G>C	Ser3749Asp Ala1929Val Asp2365His
		137	6164C>T 6039G>C	Ser2055Phe Gln2013His
24	*KIR2DL4*	119	35-4404_35-4403delGCinsCG	Ala236Arg
		137	35-12466A>G	Met118Val
25	*AHNAK2*	119	6014G>A	Arg2005Lys
		137	4105G>T	Gly1369Trp
26	*KIF25*	119	778C>A	Gln260Lys
		137	83A>T	Lys28Met

Table 4.6. Genes detected in more than one patient that had the same variant in tumour and normal adjacent tissue in the same patient, with the type of mutation and amino acid change.

4.2.8 Copy number variation detection

SNVs are the most widely studied form of genetic variation in cancers, and there are many reports of SNVs associated with risk of developing a variety of cancers, though their contributions to modification of cancer risk are generally small [56]. Besides SNVs, copy number variations (CNVs) are now considered an important form of genetic variation, and it has been reported that certain CNVs affect cancer susceptibility in individuals [56]. Recent genome-wide studies have shed light on CNVs: an unexpectedly frequent, dynamic and complex form of genetic diversity [57]. CNVs are generally identified as being inherited or de novo deletions or duplications of the genome ranging in size from 100bp to 3Mb, which may lead to changes in gene dosage and/or expression [58]. The CLC Biomedical Cancer

workbench allows us to detect CNVs in tumour sample by comparing the tumour reads to the normal tissue reads of the same patient. We have compared each of the HNC tumours to the blood reads of each patient in order to detect CNVs and study genome stability in the five patients.

The copy number variant detection tool in the CLC workbench is designed to detect CNVs from targeted re-sequencing experiments. The tool takes read mappings and target regions as input, and produces amplification and deletion annotations. The annotations are generated by a 'depth-of-coverage' method, where the target-level coverages of the case and the controls are compared in a statistical framework using a model based on selected targets.

4.2.8.1 Running the Copy Number Variant Detection tool

To start the CNV detection tool, re-sequencing analysis was chosen from the toolbox and then the copy number variant detection was selected as shown in figure 4.20.

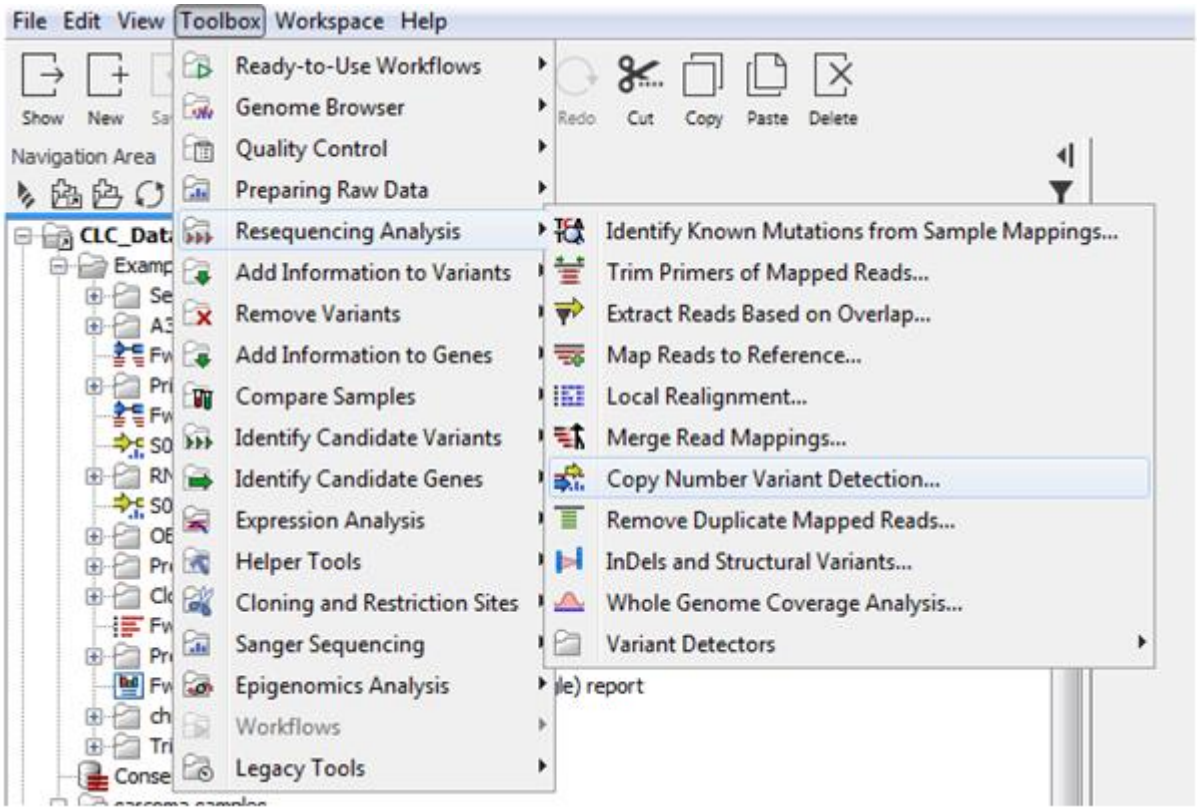


Figure 4.20. Running the Copy Number Variant Detection tool*. From the Toolbox; the Resequencing Analysis was selected. Then the Copy Number Variant Detection was selected.*

Then the case read mapping, which is the tumour read in our experiment, has been selected as shown in figure 4.21.

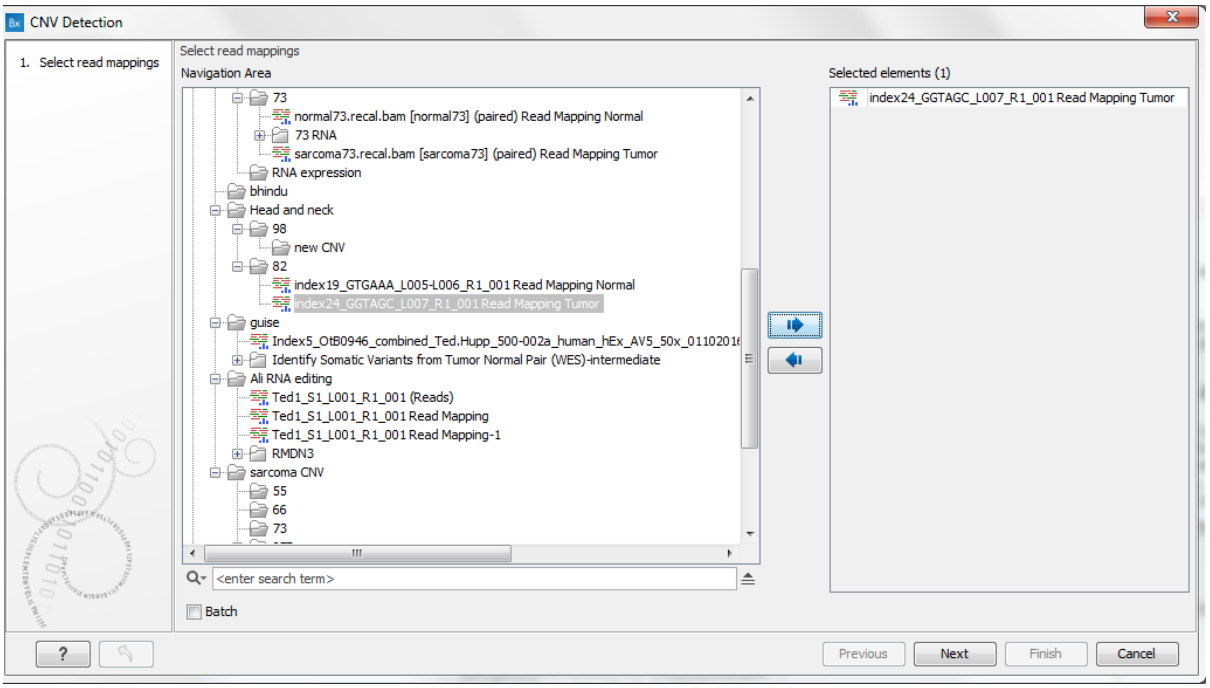


Figure 4.21. Selection of Read mappings*. In this example; the index 24 reads were selected which is the tumour reads of the patient 82.*

At the next step another browser, as shown in figure 4.22, enables us to choose the input and reference parameters. Target regions track, which is an annotation track containing the regions targeted in the experiment, was chosen. This track must not contain overlapping regions, or regions made up of several intervals, because the algorithm is designed to operate on simple genomic regions. Control mapping read was selected, which will be used to create a baseline by the algorithm. For the best results, the controls should be matched with respect to the most important experimental parameters, such as gender and technology. In our case, we have selected the normal blood reads as control read mapping. Gene track, which will be used to produce gene-level output as well as CNV-level output, was selected. Ignore non-specific matches: If checked, the algorithm will ignore any non-specifically mapped reads when counting the coverage in the targeted positions. Ignore broken pairs: If checked, the algorithm will ignore any broken paired reads when counting the coverage in the targeted positions.

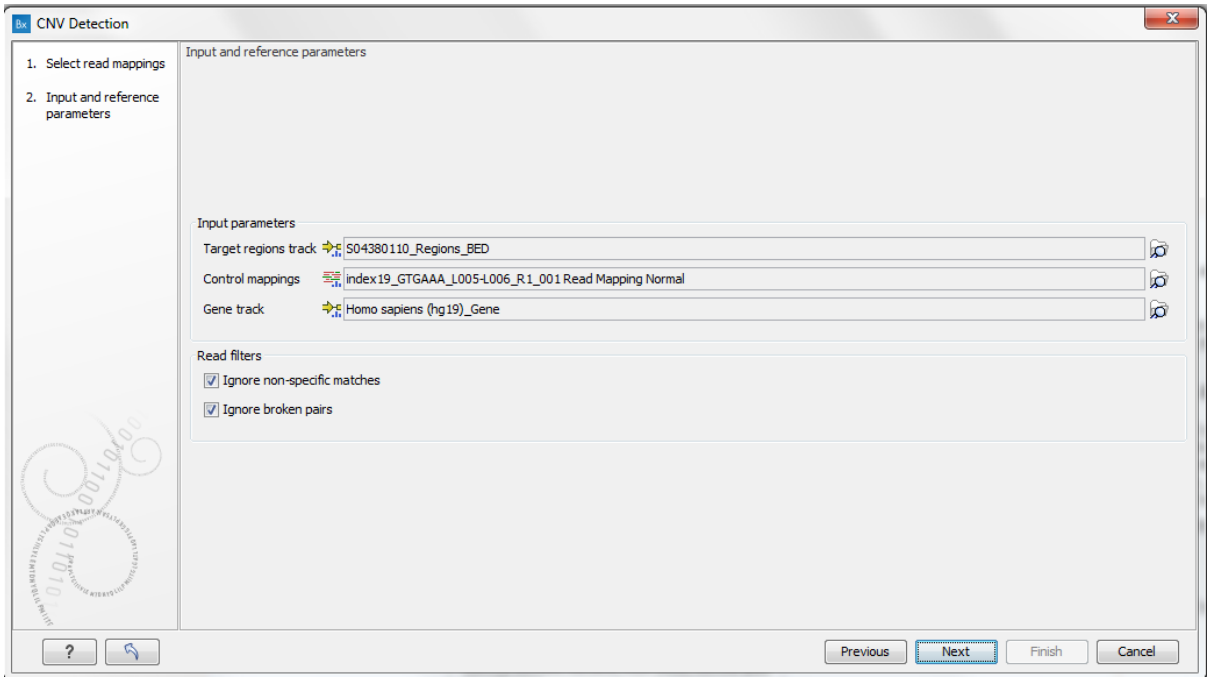


***Figure 4.22. Input and reference parameters**. Target regions track, which is an annotation track containing the regions targeted in the experiment, was chosen. Control mapping read, index 19 is the normal blood reads for patient 82, and Gene track were selected.*

Click Next to set the parameters related to the target-level and region-level CNV detection, as shown in figure 4.23. P-values lower than the threshold for significance will be considered "significant". The higher this value, the more CNVs will be predicted. In the Minimum fold change, absolute value, we had to specify the minimum fold change for a CNV call. If the absolute value of the fold change of a CNV is less than the value specified in this parameter, then the CNV will be filtered from the results. For example, if a minimum fold-change of 1.4 is chosen, then the adjusted coverage of the CNV in the case sample must be either 1.4 times higher or 1.4 times lower than the coverage in the baseline for it to pass the filtering step. For the Low coverage cut-off: if the average coverage of a target is below this value, it will be considered "low coverage" and it will not be used to set up the statistical models, and p-values will not be calculated for it in the target-level CNV prediction.

The graining level is used for the region-level CNV prediction. Coarser graining levels produce longer CNV calls and less noise, and the algorithm will run faster. However, smaller CNVs consisting of only a few targets may be missed at a coarser graining level. Intermediate: prefers CNVs consisting of an intermediate number of targets. The algorithm is most sensitive to CNVs spanning 5 or more targets. This is the recommended setting if you expect CNVs of intermediate size. Fine: prefers CNVs consisting of fewer targets. The algorithm is most

sensitive to CNVs spanning 3 or more targets. This is the recommended setting if you want to detect CNVs that span just a few targets, but the false positive rate may be increased. When finished with the settings, click Next to start the algorithm.

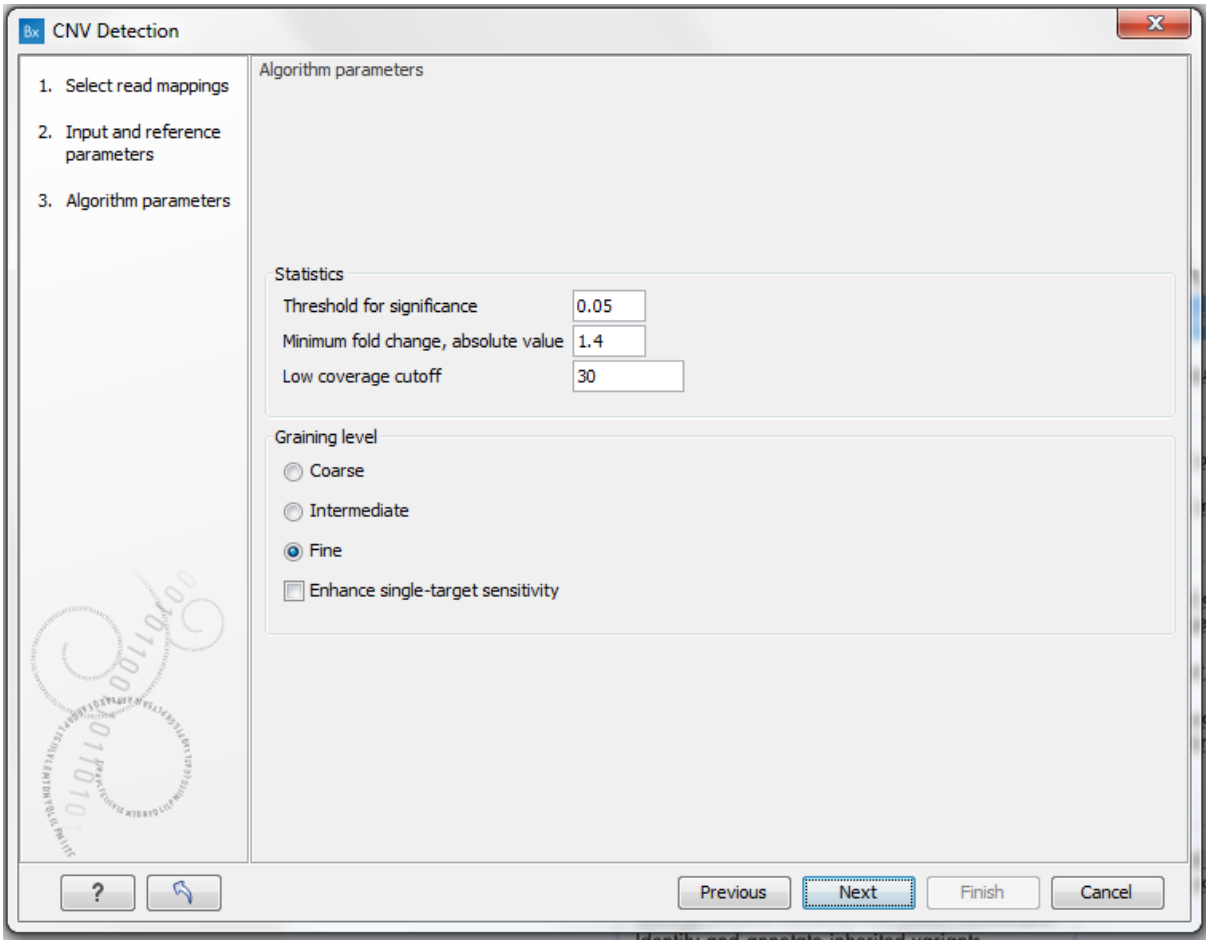


***Figure 4.23**. **The parameters related to the target-level and region-level CNV detection**. The threshold for significance was set to 0.05, so any P-value lower than the threshold for significance will be considered "significant". The minimum fold change, absolute value was set 1.4, so the adjusted coverage of the CNV in the case sample must be either 1.4 times higher or 1.4 times lower than the coverage in the baseline, for it to pass the filtering step. The low coverage cut-off was set to 30, so if the average coverage of a target is below 30 it will be considered "low coverage" and it will not be used to set up the statistical models. From the Graining level: Fine was selected which is recommended to detect small CNVs.*

4.2.8.2 CNV results

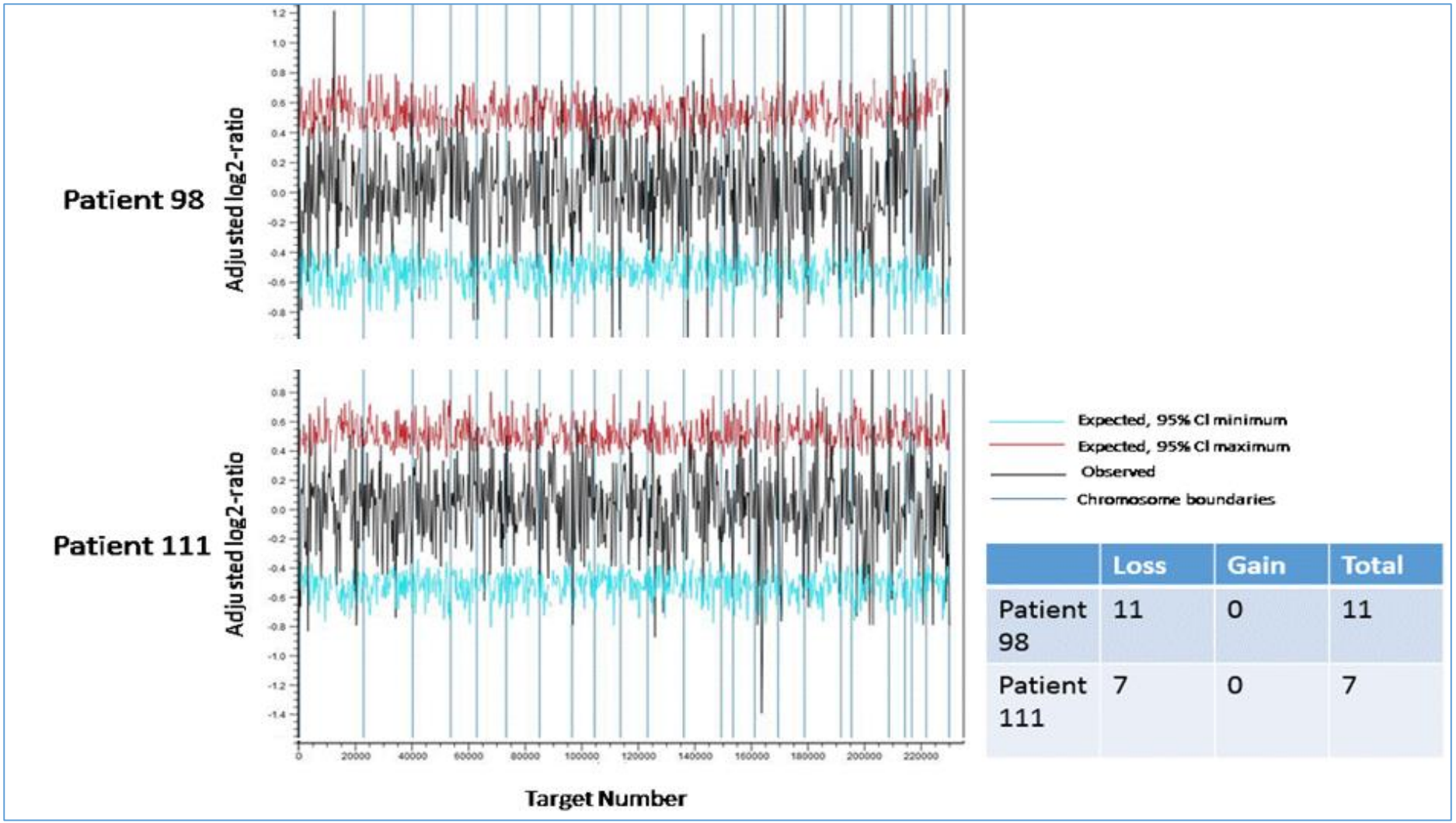
The CNV detection analysis detected small number of losses (deletions) in the young patients 98 and 111, with 11 and 7 deletions respectively. There is no gain detected in both patients (fig 4.24). There is no common CNV in the two patients.

There are more CNVs detected in the young patient 82 and the old patients 119 and 137 than the young patients 98 and 111. There are 27 losses and 27 gains in patient 82; 33 losses and 5 gains in patient 119; and 13 losses and 11 gains in patient 137 (fig 4.25). There is no common CNV (exact gain or loss region) in these three patients. However, there are some overlapping regions of gains or losses between them: patient 82 shares an amplified region on chromosome 9 with patient 119, as shown in figure 4.25. Table 4.7 shows genes that are in these overlapping regions.

The young patient 82 and the old patients, 119 and 137, all have p53 mutations, whereas in the young patients 98 and 111 there are no mutations detected in p53. This may explain why the old patients and patient 82 have more genomic instability and more CNVs than patients 98 and 111.

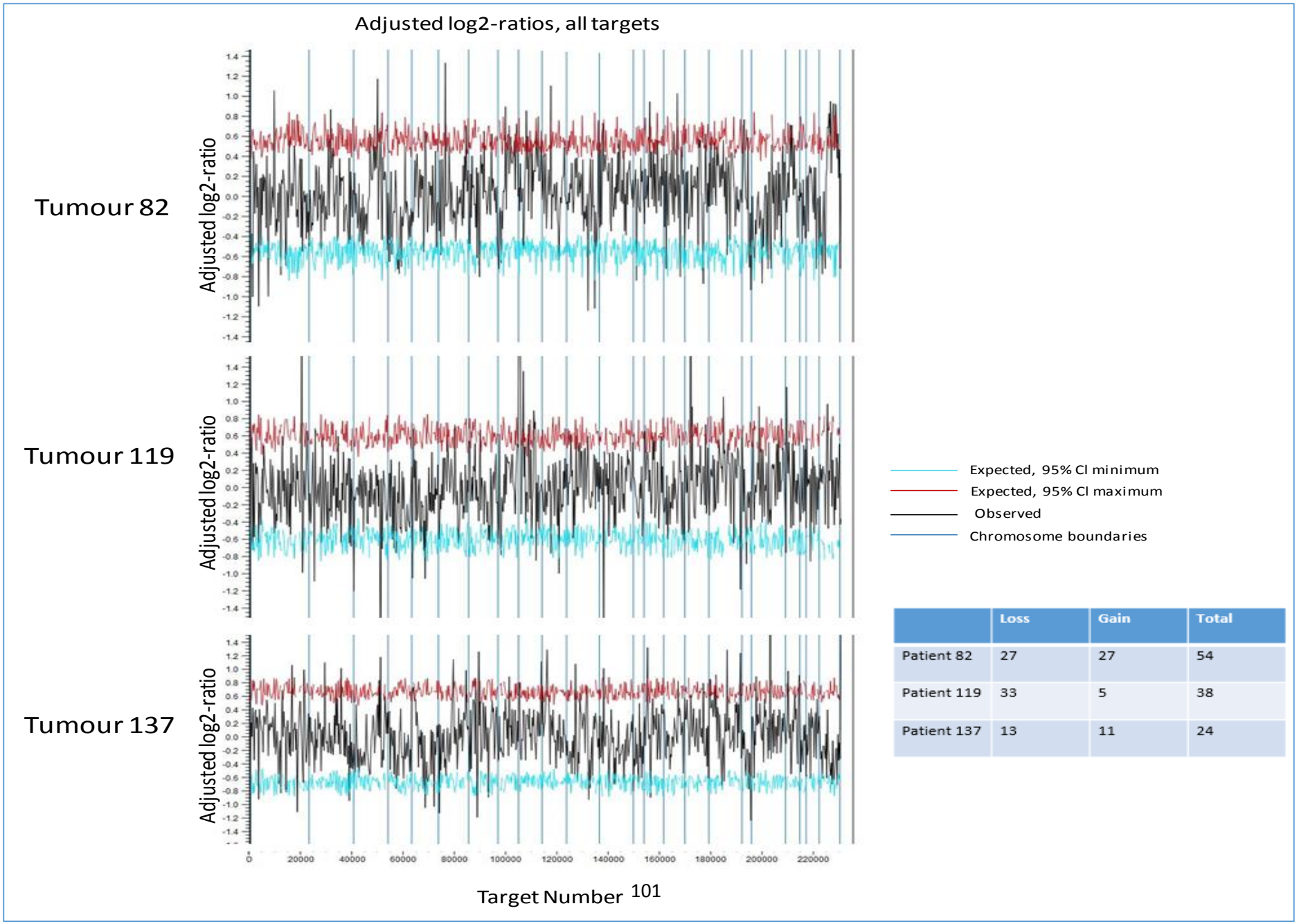
Number	Chromosome	Gene	Patients	Consequence
1	9	*RCL1*	82, 119	Gain
2	9	*MIR101-2*	82, 119	Gain
3	9	*RP11-125K10.5*	82, 119	Gain
4	9	*JAK2*	82, 119	Gain
5	9	*AL161450.1*	82, 119	Gain
6	9	*INSL6*	82, 119	Gain
7	9	*INSL4*	82, 119	Gain
8	9	*RLN2*	82, 119	Gain
9	9	*RLN1*	82, 119	Gain
10	9	*PLGRKT*	82, 119	Gain
11	9	*CD274*	82, 119	Gain
12	9	*PDCD1LG2*	82, 119	Gain
13	9	*KIAA1432*	82, 119	Gain
14	6	*DEFB114*	82, 137	Gain
15	6	*DEFB113*	82, 137	Gain
16	8	*EFCAB1*	82, 137	Gain
17	8	*SNAI2*	82, 137	Gain
18	8	*C8orf22*	82, 137	Gain
19	8	*RP11-738G5.2*	82, 137	Gain
20	8	*SNTG1*	82, 137	Gain
21	18	*CCDC178*	119, 137	Loss

Table 4.7. List of genes in the shared CNVs regions in patients 82, 119 and 137.



	Loss	Gain	Total
Patient 98	11	0	11
Patient 111	7	0	7

Figure 4.24. A graph showing the mean adjusted log-ratios of coverages in the report produced by the Copy Number Variant Detection tool for patients 98 and 111 by comparing tumour reads of each patient to the normal blood reads of the same patient*. A table beside the graph shows the number of CNVs detected in each patient. The CNVs are detected when the log-ratios of coverages of targets on the chromosomes are significantly higher or lower than for targets on other chromosomes. The black line in these regions is outside the boundaries defined the cyan and red lines.*



	Loss	Gain	Total
Patient 82	27	27	54
Patient 119	33	5	38
Patient 137	13	11	24

Figure 4.25. A graph showing the mean adjusted log-ratios of coverages in the report produced by the Copy Number Variant Detection tool for patients 82, 119, and 137 by comparing tumour reads of each patient to the normal blood reads of the same patient*. A table beside the graph shows the number of CNVs detected in each patient. The CNVs are detected when the log-ratios of coverages of targets on the chromosomes are significantly higher or lower than for targets on other chromosomes. The black line in these regions is outside the boundaries defined the cyan and red lines*

4.3 Discussion

4.3.1 Setting the parameters of the variant

I have learned to set the parameters to detect variants that appear in at least two DNA sequencing reads in the tumour files, and at least 5 DNA sequencing reads (coverage) at the mutation site, to make sure it is a real variant and not a sequencing error variant. For the normal reads we have chosen 1 read at the beginning of the analysis, which resulted in many detected variants. DMKN gene was one of the genes detected in all five patients. It has a 51 base deletion in the tumours of patients 98 and 111, and a 12 base insertion in patient 137, with the same deletion seen in the control (blood), but with a very low coverage of 4, 1, and 2 respectively. We have designed primers to amplify the region that has the deletion and insertion to validate these indels. The PCR results (Fig 4.26) showed that the deletion found in tumours from patients 98 and 111 is also found in blood and in tumour adjacent tissue, as shown in figure 3.26 below, which means that it is a false positive variant, although it is reported in COSMIC as a somatic mutation (COSM69118), along with the insertion in patient 137. For this reason, we have selected the minimum coverage in the control to be at least 5.

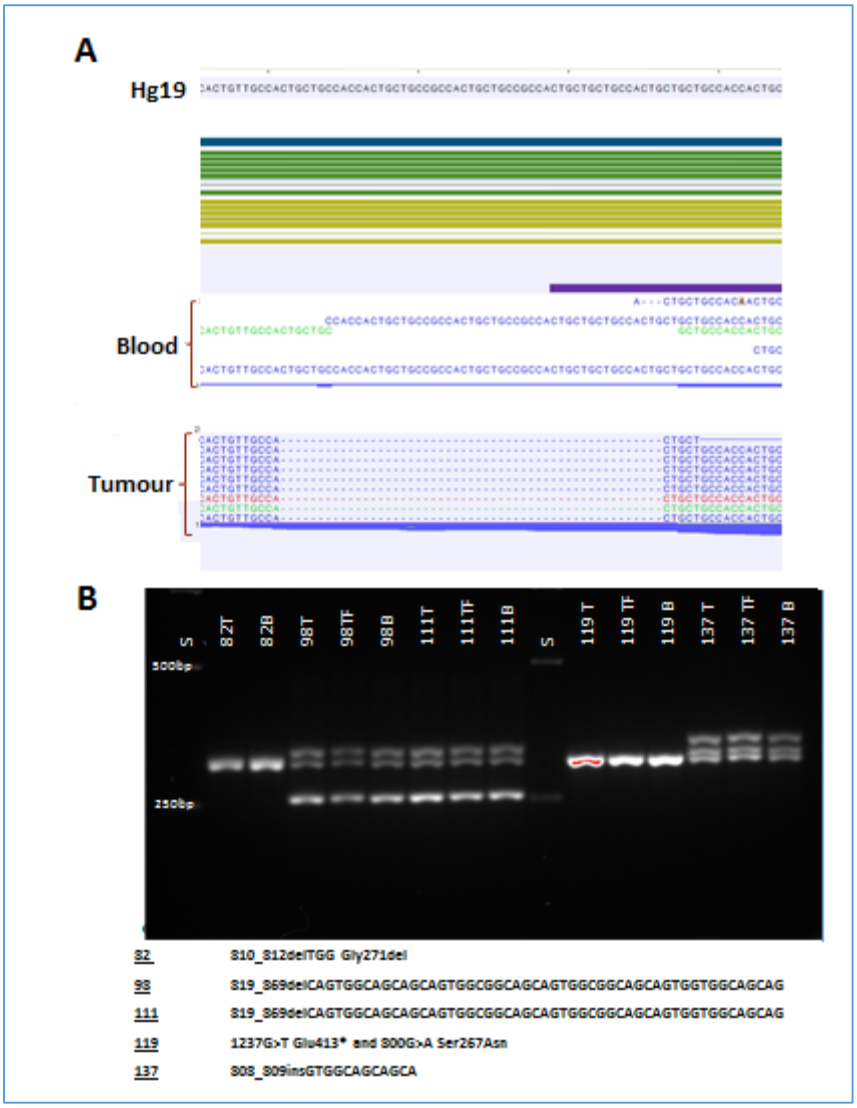


***Figure 4.26. Validation of detected mutations in DMKN gene by PCR**. **A**, CLC browser shows the deletion in patient 111. **B**, PCR result shows the deletion in 98 and 111 patients in tumour (T), normal adjacent tissue (TF) and blood (B). The detected insertion in patient 137 is also in T, TF, and B.*

4.3.2 Factors affecting the number of mutations and mutation signature

In young patients, the number of non-synonymous mutations is similar in the female patients 82 and 111, whereas the male patient 98 has more than double the number found in each of the female patients. The older male patient 119 has a higher number of mutations than the female patient 137. As all of the patients in this study have no history of smoking or alcohol consumption and are negative for HPV infection the reason for men to have higher number of mutations than women is unknown. The old patients have a higher number of mutations than the young patients; this is not surprising, as the number of accumulated mutations is known to increase with age [59].

All patients have the same types of mutation. They all show the highest frequency of C>A with C>T is the second highest. The C>A signature is observed in lung adenocarcinoma, squamous and small cell carcinoma, head and neck squamous, and liver cancer, most of which are believed to be caused by tobacco smoking [26]. All the patients in this study have no smoking history, so there must be another environmental or biological carcinogen causing this type of mutation in both young and old patients. Pickering, C. R., et al. 2014, found that the types of base changes observed in squamous cell carcinoma of the oral tongue (SCCOT) in young patients were similar to those observed in old patients despite the difference between the two groups in smoking history. They also found that smoking has a minor impact on the types of mutations in SCCOT. It was shown that the C>A type is common in hepatocellular cancers and this is believed to be associated with aflatoxin, a known carcinogen commonly found in food from South Africa and Asia [27].

All patients in this study were negative for HPV infection. The young age of the patients with no carcinogen exposure raises the possibility that a virus, either known or novel, could be the reason for cancer development [60]. However, Li, R., et al. (2015) failed to identify any potentially causative viruses, including HPV, in 19 tongue tumour samples through transcriptomic analysis using 3 separate approaches. They therefore suggested the “hit and run” hypothesis, which states that the oncogenic viruses can either integrate into the host cell genome or remain episomal at a particular point in the progression to cancer. The viral genome may be subsequently lost and become undetectable at the time of diagnosis.

There is a diverse population of microorganisms in the oral cavity. Some of these microorganisms could also play a role in carcinogenesis. A number of microorganisms have already been associated with the development of oral cancer including species of *Streptococcus* and *Candida* [59].

4.3.3 Common mutated genes detected

The CLC biomedical workbench analysis detected a very large number of genes with non-synonymous mutations in each patient. This number was lowered to 109 genes after the cut-off of 15% frequency (count/coverage) applied and specification that the gene must be detected in at least two patients. It is challenging to find out which of these mutated genes are driver genes. A driver gene is a gene which, when mutated, is responsible for the initiation or progression of a cancer, or a gene with a mutation that occurs more often than expected by chance [61]. Many of the mutations detected in the cancer genome, called passenger mutations, have no effect on the development of the cancer. These are attributed to the inherent instability of the cancer genome [61]. Beyond statistical significance, there are other factors to be considered when selecting driver mutations. These factors include whether the mutation is at a specific functional position of the gene, and whether the gene has been reported in any other cancer.

Four genes out the 109 genes listed above; have been reported before as commonly mutated genes in HNC. *TP53, CDKN2A* and *FLG* were reported in head and neck squamous cell carcinoma [61]. *HERC2* was reported in carcinoma of the oral tongue (OTSCC) [62].

When the genes in table 4.3 (109 genes with ≥ 15% frequency) were examined using the DAVID gene functional classification tool to see which genes were affected in different cancer pathways, five genes appeared in the chart as shown in figure 4.27. These genes were *TP53, CASP9, CTBP2, MSH3*, and p14 ARF and p16 INK4a that are encoded by *CDKN2A* gene.

4.3.3.1 *TP53* gene

TP53 has been detected in three patients: 82 with three mutations, 119 with one mutation and 137 with two mutations (see additional file). *TP53* is the most commonly mutated gene in head and neck cancers and is detected in approximately 50% of cases; this rate increases in HPV-negative cases to 70–80% [63]. HNC appears to be the most common p53 mutation-carrying cancer type after ovarian cancer and lung squamous carcinoma [64]. *TP53* mutations can occur throughout the entire gene but the majority of these mutations are missense mutations, causing single amino-acid substitutions in the DNA binding domain encoded by exons 5 to 9 [44, 63]. The gene codes for a tumour suppressor protein (p53) which plays a key role in the regulation of genes involved in the cell cycle and growth arrest, DNA repair and apoptosis (programmed cell death), thereby maintaining genomic stability [64]. When the

function of p53 is lost, these protective mechanisms are disabled, which leads to a higher propensity for the cell to accumulate and propagate genomic insults [63]. The p53 level is regulated by mouse double minute 2 (*MDM2*), an E3 ubiquitin protein ligase that binds to p53 and causes its degradation [64]. MDM2 is inhibited by p14ARF, which is encoded by the *CDKN2A* gene, protecting p53 from degradation [64]. There is no therapeutic strategy being used to target tumours with p53 mutations because p53 is a tumour suppressor protein. However, some studies have examined delivery of functional p53 via gene therapy using an adenoviral vector [63].

4.3.3.2 *CDKN2A* gene

The *CDKN2A* gene was detected in three patients: 82 and 137 with high frequency (41.63% and 32.29% respectively), and in 98 with low frequency (5.26%). *CDKN2A* encodes the proteins p14 ARF and p16 INK4a, which are generated by alternative splicing. The role of p14 ARF is the protection of p53 as mentioned above. The p16 INK4a antagonises cyclin dependent kinases 4 and 6 (CDK4/6), resulting in blocking phosphorylation of the retinoblastoma protein pRb and, consequently, cell cycle progression [65]. Consistent with the CDKN2A role in the regulation of the p53 and pRb signalling pathways, which are required for the induction of apoptosis in response to a number of cellular stressors such as DNA damage, mutations in this gene are prevalent in cancers and play an important role in oncogenesis and tumour progression [65].

4.3.3.3 *CASP9* gene

CASP9 was detected in two patients; 82 and 137. CASP9 has never been reported in HNC before, but other caspases have been reported, such as *CASP8* [61]. Caspases are crucial for apoptosis, so their inactivation can lead to the persistence of mutated cells and promotion of tumorigenesis [66]. The reduced expression of proapoptotic caspases has been reported in different cancers, and specific mutations have been linked to various tumour types and stages of transformation [66]. It has been reported by [67] that the use of a specific inhibitor of CASP9 almost completely blocked apoptosis induced by cisplatin – one of the most effective anticancer drugs widely used for the treatment of different tumour types including HNC.

4.3.3.4 *MSH3* gene

The *MSH3* gene was detected in the two old patients: 119 and 137. It belongs to the DNA mismatch repair system family of proteins whose function is to correct base-base mispairs, introduced into the DNA during DNA synthesis, to maintain genomic stability [68]. It has been reported that MSH3-dificient mice develop late onset microsatellite instability-positive gastrointestinal cancer, suggesting that this deficiency can contribute to tumour initiation [68].

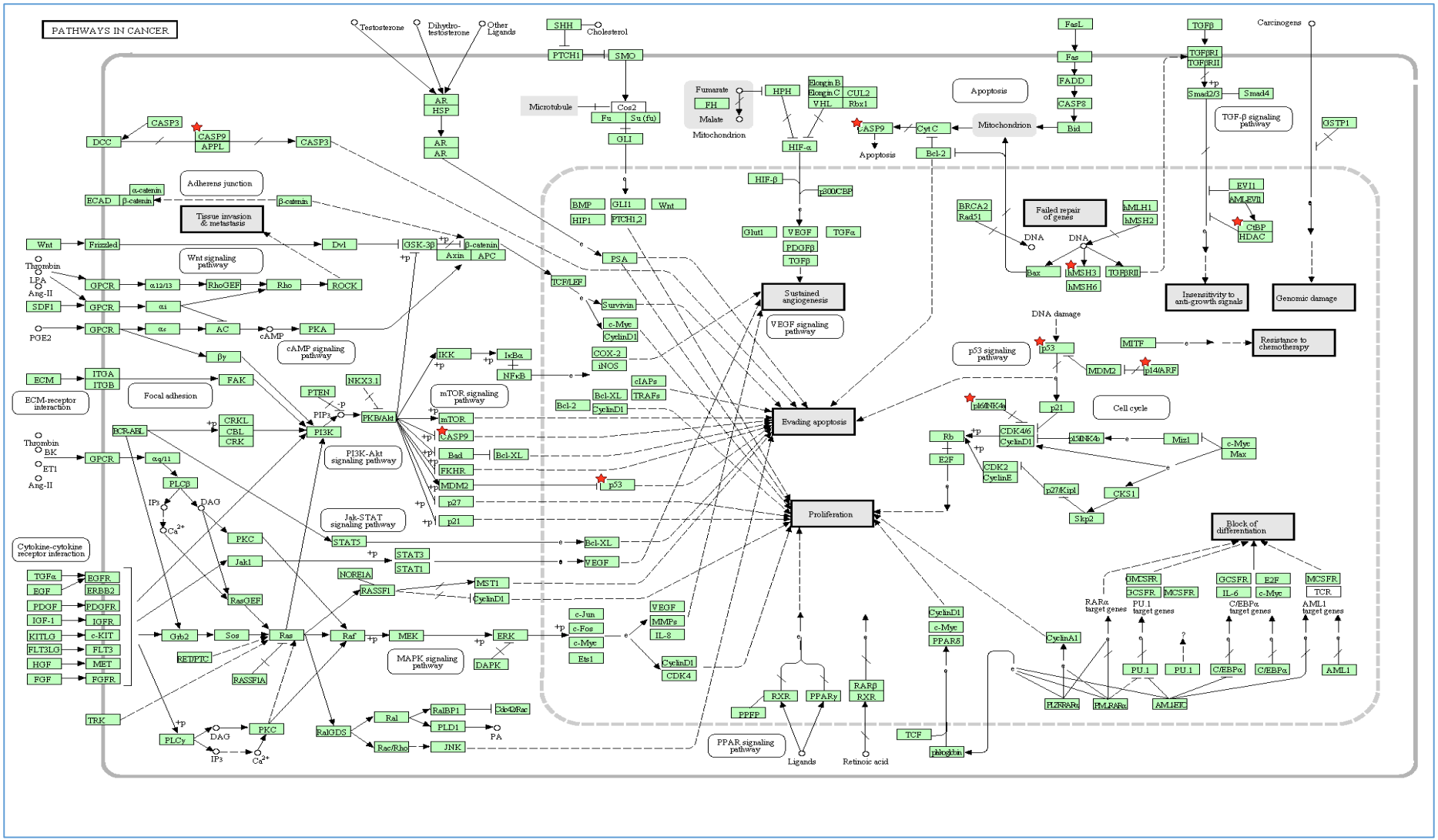


***Figure 4.27*. *A chart of the different genes involved in different cancer pathways made by the DAVID gene functional classification tool*.** *The genes from our study involved in these pathways are marked by a red star.*

4.3.3.5 *CTBP2* gene

The mammalian C-terminal binding protein 2 (*CTBP2*) gene has been detected here in HNC for the first time. CTBP2 promotes cancer migration, invasion, and survival by acting as a transcriptional co-repressor of multiple tumour suppressor genes involved in pathways associated with tumourigenesis, including TGF-β and Wnt signalling pathways, and cell cycle regulators such as RB and MDM2 [69, 70]. Increased expression of CTBP2 was reported in tumours from patients with head and neck squamous cell cancers [71]. CTBP2 was also found overexpressed in tumours of colon, breast, and prostate, and this high expression was found to be associated with a poor prognosis in these patients [69, 72].

CTBP2 plays an important role in stimulation of cell migration by the RAC pathway via regulation of T-cell lymphoma invasion and metastasis 1 (Tiam1) protein, which is a guanine nucleotide exchanger factor (GEF) for RAC GTPase that plays a critical role in regulating cell adhesion, invasion, and migration, and has been directly implicated in the promotion of cancer progression and metastasis [69]. Tiam1 was detected as a transcriptional activation target of CTBP2 and the overexpression of CTBP2 increased Tiam1 expression, leading to increase cell migration [69].

It was reported by [73] that MDM2 represses p53 activity through the recruitment of CTBP2. They showed that this interaction between the MDM2 and CTBP2 has much greater effect on p53 repression.

CTBP2 appears also in the Notch signalling pathway as shown in figure 4.28 below. The Notch1 gene was reported to be mutated in 12%–15% of tumours of HNSCC, and mutations were detected in other Notch family members in 3%–5% cases [74]. It was reported that reduced activity of the Notch signalling pathway is associated with the development of chronic myelomonocytic leukaemia, suggesting a tumour suppressor role for Notch [47].

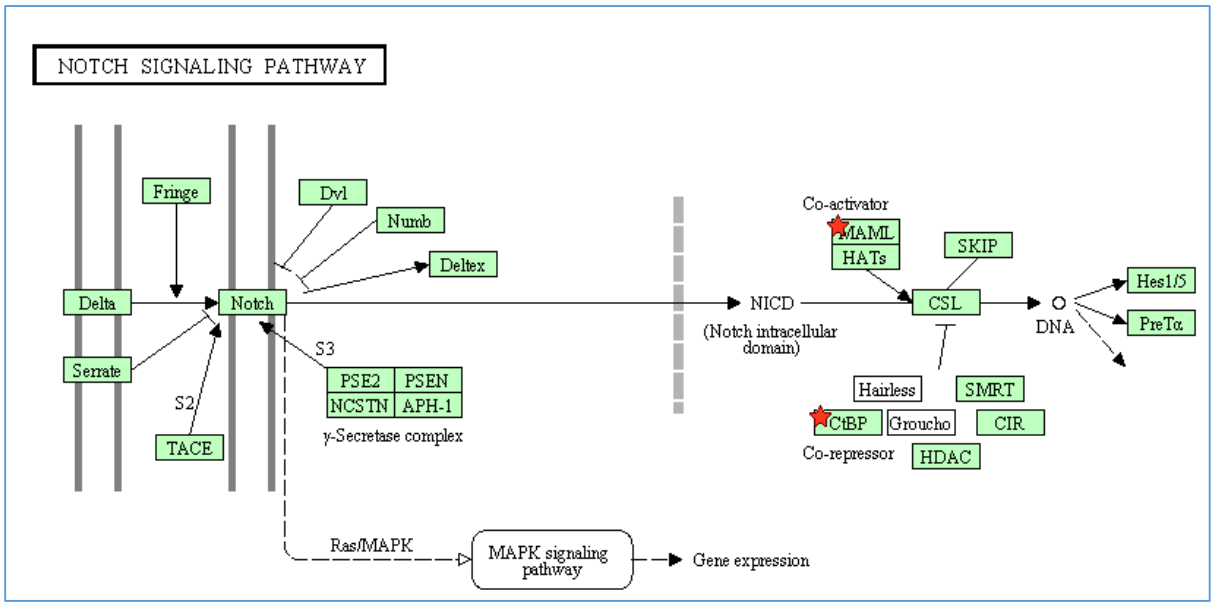


***Figure 4.28**. **A chart of genes involved in the NOTCH signalling pathway by the DAVID gene functional classification tool**. Genes from our study are marked by red star*.

4.3.4 Field cancerization

The normal adjacent tissue had many more mutations than the tumour in all patients, and there were small number of common mutations between the tumour and the normal adjacent tissue, which is compatible with the field cancerization theory. The process of carcinogenesis begins with a stem cell, which develops one or more genetic alterations. Subsequently a clone of genetically altered cells forms a patch. Figure 4.29 below shows the steps of the field development. The presence of a field with genetically altered cells appears to be a continuous risk factor for cancer as cancer can develop from such a field that is left behind in the patient after surgery [75]. The oral cavity was proven to be most susceptible to field formation, as it is exposed to a wide range of environmental carcinogens which affect the entire mucosa and result into the simultaneous occurrence of premalignant states [76]. A field size of over 7 cm has been reported in HNC, and about 62.5% of HNSCC second primary tumour recurrences are from similar clonal fields left behind after resection [77]. Therefore, a frequent oral examination with histological studies and molecular testing are important for patients after surgery, especially for those at high risk of developing malignancies.

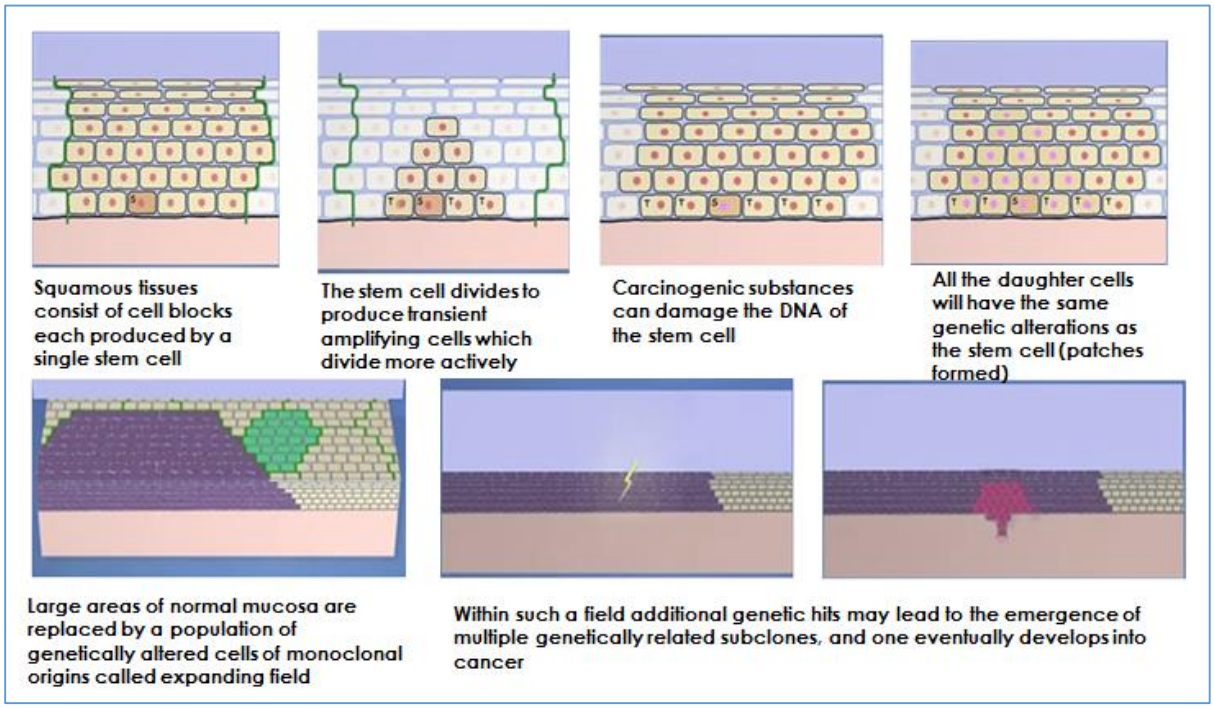


Figure 4.29. The steps of field cancerization formation.

4.3.5 Copy number variation detection

The genomes of the old patients and the young patient 82 are more unstable, as they harbour more CNVs than the young patients 98 and 111. This may be explained by the presence of p53 mutations in patient 82 and old patients, which increases genome instability. Some mutations in p53 have been shown to promote aneuploidy and tumorigenesis in the mammary gland in transgenic mice and to affect genomic stability, in part, by causing centrosome abnormalities [78]. The high number of detected CNVs in patient 82 could explain the highly aggressive tumour in this patient (the patient died shortly after diagnosis).

4.4 Conclusion

The SNV and CNV analysis showed that there were two different genetic groups in the five patients, independent of the age of the patients. The first group has mutations in the p53 and have more CNVs, which is seen in the old patients 119 and 137, and in the young patient 82. The other group does not have mutations in p53 and has very few CNVs.

In order to detect a specific tumour mutation that can be used as a tumour target, it is necessary to compare the tumour sequences with the normal adjacent tissue rather than with blood to avoid selecting shared mutations between tumour and the adjacent tissue. This has important implications for identifying truly tumour specific mutations, since the vast majority of cancer genomics studies uses normal blood as the source of germline mutations, but does not usually use normal adjacent tissue as the source of "normal" DNA. Ideally, all three samples types would be taken, but this is sometimes limited by clinical practise (such as in my study on UPS). Nevertheless, the high mutation rate in normal tissue, some of which mutations are shared with the tumour, highlight the concept that mutations are not necessarily cancer causing, but that clusters of mutations are required to produce the transformed cell.

CHAPTER FIVE
Analysis of somatic mutations in 20 undifferentiated pleomorphic sarcoma

5.1 Introduction

5.1.1 Sarcoma Epidemiology

Sarcomas are uncommon, diverse mesenchymal malignancies that arise in, or from, bone, cartilage or connective tissues, such as muscle, fat, peripheral nerves and fibres [79]. They represent approximately 1% of cancers diagnosed in adults and 15% of childhood tumours [80]. Regardless of their tissue of origin, sarcomas share an overall poor prognosis. Because they are frequently discovered at more advanced stages and there is limited understanding of their pathogenesis, there is little guidance available for developing targeted therapies [81].

Sarcomas show remarkable histological diversity, with more than 50 recognized subtypes [82]. Because of the heterogeneity of sarcoma tumours, their true incidence has generally been underreported. Recent estimates from the cancer networks suggest that about 3000 patients are diagnosed each year in the UK [83]. Approximately half of all sarcoma patients with intermediate or high-grade tumours develop metastatic disease. The overall survival rate is about 50% at 5 years [83].

From a molecular perspective, sarcomas are classified into two categories. The first category comprises sarcomas with near-diploid karyotypes and simple genetic alterations, including translocations or specific activating mutations. The second category includes tumours with complex and unbalanced karyotypes, such as the undifferentiated high-grade pleomorphic sarcoma (UPS) [79, 80]. UPS is the most common form of adult sarcoma, forming aggressive tumours, which frequently show local recurrence and can metastasise to distant sites [80].

5.1.2 Sarcoma aetiology

The aetiology of most sarcomas is unknown, but in a minority of sarcoma patients an association exists with irradiation, chronic lymph angioedema, viral infections and known genetic susceptibility [84]. There is an increased risk of sarcomas, both bone and soft tissue, in patients who have had a familial retinoblastoma caused by inherited mutations in the *RB* gene. Also, there is an increased risk of sarcomas, and other cancers, in families with Li–Fraumeni syndrome who have inherited mutations in the *TP53* gene [83].

5.1.3 Sarcoma management

Sarcoma tumours have traditionally been managed by wide excisional surgery with or without radiotherapy. The use of chemotherapy or targeted therapy has mostly been reserved for advanced disease, aiming to achieve disease palliation and control [84].

Due to the availability of increasing amounts of molecular data from expression profiling and other molecular techniques, differential diagnosis and subgrouping of sarcomas has been facilitated. Novel targeted treatment regimens should be developed on the basis of these results. High-throughput sequencing is of great help in identifying novel targets and is further supported by new applications such as RNA sequencing [85].

5.1.4 Molecular mechanisms of sarcoma

The mechanisms that drive human sarcomagenesis fall into three broad categories: transcriptional dysregulation owing to aberrant fusion proteins that result from genomic rearrangements, somatic mutations in key genes and signalling pathways, and DNA copy-number abnormalities [79].

The improvements in molecular techniques and the expanded use of NGS have allowed identification of targetable genetic alterations in specific cancer types.

Gastrointestinal stromal tumour (GIST), one of the more common types of human sarcoma, is characterized by oncogenic mutations in KIT, and in platelet-derived growth factor receptor-α (PDGFRA). The dependence of GIST on constitutively activated KIT and PDGFRA has led to treatment with the selective kinase inhibitor imatinib, which achieves a partial response or stabilises the disease in about 80% of patients with advanced or metastatic GIST [79].

5.1.5 Exome sequencing studies to identify mutations in sarcoma

New research has analysed targeted exon sequencing, which was done for 72 genes (selected because of their associations with increased cancer risk) of 1162 patients with sarcoma and 6545 Caucasian controls, in order to investigate the genetic basis of sarcoma [86]. The study reported that 638 (55%) of the probands bore an excess of pathogenic germline variants. Therefore, about half of sarcoma patients have putatively pathogenic variations in known and novel cancer genes. This may explain the development of sarcoma in children and young-age patients. This study focused on 72 genes, and focused on detection of germline mutations. There might be other genes with mutations specific to sarcoma. In order to develop a drug target, a somatic mutation that is specific to a tumour must be detected. Therefore, we have analysed whole exome sequencing of tumour–normal pairs of 20 UPS patients, in order to identify tumour specific mutations.

5.1.6 Analysis of 20 UPS whole exome sequences to identify somatic mutations

The widespread use of NGS has uncovered the genomic landscape in many tumour types. Given the disease rarity and histologic diversity, much less is known about the somatic mutational landscape in sarcoma.

UPS is defined as a sarcoma with no identifiable line of differentiation, excluding dedifferentiated types of specific sarcomas [87]. Biomarkers that can distinguish UPS from other types of sarcoma, as well as improve treatment stratification, are needed. Previous genetic analysis studies have failed to reveal any consistent or tumour specific aberrations [87].

We have used the CLC Biomedical Genomic Workbench to analyse the NGS of 20 UPS exomes and their matched normal exomes to define somatic mutations, including single nucleotide variants (SNV), small insertions or deletions (INDELS) and copy number variants (CNVs), which may lead to identification of novel genomic alterations that could serve as therapeutic targets.

5.2 Results and Discussion

Tumours were sequenced to an average depth of 100x coverage and matched DNA samples from morphologically normal tissue adjacent to the tumour were sequenced to 30x. Exome sequencing was performed using Agilent V5+UTR Exome Capture Kit (75Mb); Illumina, 100bp paired-end reads using a coverage of tumour–normal pairs (100X/30X). Paired de-multiplexed fastq files were generated using CASAVA software (Illumina). Somatic mutations were identified using the CLCbio Genomics Workbench.

5.2.1 Identification of somatic variants from UPS tumour–normal pairs

We have used the CLC Biomedical Workbench 2.5 to analyse the NGS of 20 UPS exomes and their matched normal exomes to define somatic mutations including SNVs, Indels and CNVs. The parameters chosen for the minimum coverage, minimum count and minimum frequency of the detected variant were 5, 2 and 5 respectively in the tumour reads. We selected to remove any variant that presents at least once in the normal reads, so that all the detected variants are tumour specific. The minimum coverage for the variant site has been set to 5 in the normal reads, so any variant with less than 5 reads in its site in the normal reads would not be called out.

As described in the previous chapter, the 'identify somatic variants from tumour–normal pair' (WES) ready-to-use workflow was used to identify potential somatic variants in a tumour sample, when compared to the control sequences of the same patient.

5.2.2 Number of non-synonymous variants detected

There were many variants detected in each patient, but we were interested in non-synonymous variants only. The number of non-synonymous variants ranges from 235 to 612, as shown in table 5.1 below, with SNVs being the most detectable type of variants in each patient.

No.	Patient	No. of Variants detected by CLC	No. of SNVs detected by CLC
1	52	526	412
2	55	516	399
3	59	346	242
4	60	476	359
5	66	434	293
6	73	399	274
7	74	574	506
8	84	539	364
9	90	490	312
10	94	612	434
11	97A	379	281
12	97B	472	346
13	197	316	207
14	244	318	237
15	258	284	195
16	297	402	297
17	343	449	311
18	364	334	233
19	378	349	239
20	430	602	423
21	496	235	152

Table 5.1. Summary of the non-synonymous mutations detected in UPS*. The data summarise the number of total variants (including in-frame deletions or insertions, as well as single nucleotide variants (SNVs)) and the subset that are only SNVs. Numbers 11 and 12 represent a tumour from patient 97, in which the tumour was divided into two parts to establish intra-tumour heterogeneity. For example, the number of variants or SNVs in 97a vs 97 b was 379 vs 472 and 281 vs 336, respectively.*

5.2.3 Validation of some of the detected mutations by Sanger sequencing

We have picked up three mutations (table 5.2) to validate by Sanger sequencing. Two of these mutations are in patient 55; both have high frequency. One of them has high control coverage and one has low control coverage. The third mutation is in patient 73, and has with relatively low frequency and high control coverage. As shown in figure 5.1 below, the Sanger sequence has confirmed the presence of the detected mutations in all of the three genes. However, the mutation in the *Sema6a* gene, which has low control coverage, is a false positive as the normal

sequence shows that the normal tissue is heterozygous for this variant. This germline mutation failed to be detected in the normal genome, most probably due to the low coverage at the mutation site in the normal genome. At the same time, it was successfully detected in the tumour genome by the analysis software which misrepresented the mutation as a tumour-specific mutation. Coverage at any given place in the genome is variable however, and low coverage regions may still lead to false-positives when coverage is not high enough to identify germline variants [88].

The effect of tumour–normal coverage ratios on variant calling was investigated to assess whether increasing coverage of the tumour alone is sufficient to increase mutation detection sensitivity [89]. The 250 coverage tumour genome was compared with control genomes at 250, 200, 150, 100, 50 and 30 coverages. Down to 150 coverage, few differences were seen in the mutations called when compared with 250–250 tumour control coverage. At lower control coverage levels, a notable increase is observed in the overall number of mutations reported, because of a sharp rise in those called with a low allele fraction which appeared to be somatic when the control coverage is insufficient to show the same phenomenon [89]. Thus, keeping the tumour: normal ratio of coverage closer to one appears to play a role in maintaining the accuracy of mutation calling.

Number	Patient	Gene Name	Mutation	Count	Coverage	Frequency	Control Coverage
1	55	Il11ra	649C>T	143	158	90.51%	55
2	55	Sema6a	947G>T	55	60	91.67%	14
3	73	Fat3	2855G>A	70	260	26.92%	61

Table 5.2. Three mutations were chosen for Sanger sequence validation*. The table highlights the patient number, gene name, mutation site, mutation count, tumour mutation site coverage, and frequency of mutant reads/normal reads in the tumour and the control normal tissue sequencing coverage.*

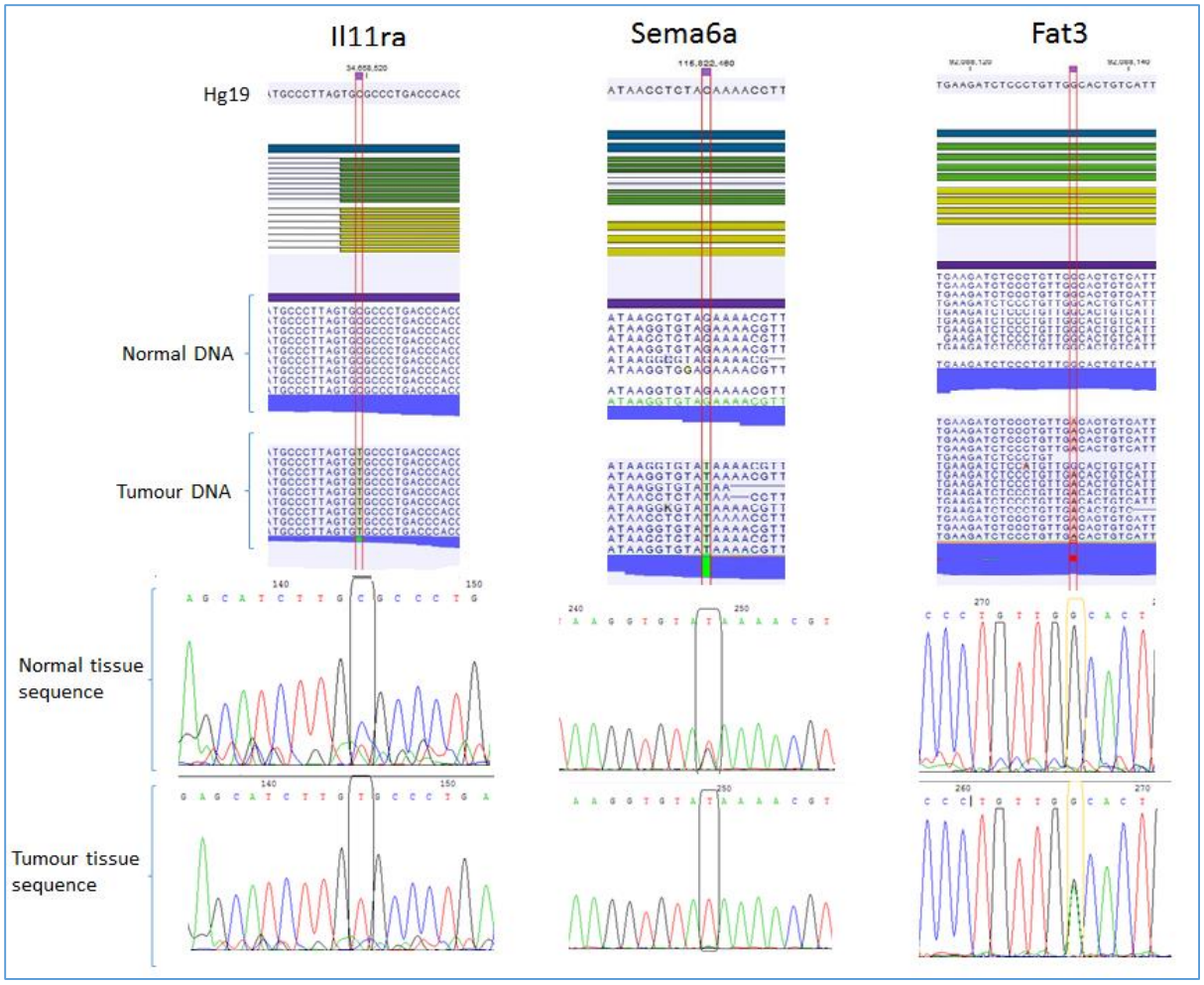


Figure 5.1. ***Validation of the three detected mutations by Sanger sequences***. *The browser highlights the hg19 reference sequence, the reads in normal DNA, the reads in tumour DNA (highlighting mutation in colour), and the Sanger sequencing validation. As an example of positive validation, Il11ra (C>T) and Fat3 (G>A) have been validated as somatic mutations by Sanger sequencing. As an example of negative validation, the mutation in Sema6a (G>T) was detected in the normal adjacent tissue, although it is possible that tumour cells contaminated the normal adjacent tissue. Nevertheless, we find that when the normal tissue coverage is low (as in SEMA6a which has only 14 reads in normal tissue) that the false positive rate is high.*

5.2.4 Common mutated genes in UPS patients and their pathways

To our knowledge, this is the first study to detect somatic mutations in UPS (and sarcoma in general) using analysis of whole exome sequences.

To look for common mutated genes in the 20 sarcoma patients, we have selected genes with mutations of ≥ 15% frequency. There were many genes with somatic mutations detected in ≥ 4 patients as shown in figure 5.2. These genes have different mutations of SNVs and Indels.

TP53 is one of the detected genes, which is mutated in 20% of samples, with SNV and a 15-base deletion in patient 52, SNV in patients 55 and 74, and a one-base deletion in patient 94.

TMPRSS13 (Transmembrane Protease, Serine 13) gene is another example, which was detected in 9 patients. Table 5.3 below shows the types of mutations detected in this gene in the different patients, with the count, coverage and frequency of these mutations in the tumours, and the control tissue coverage. All the mutations have high frequency and high control coverage. There are five patients that share the same in-frame deletion of 15 (194–208) bases. There is another in-frame deletion of 15 (248–262) bases detected in three other patients.

TMPRSS13 is a member of the hepsin/TMPRSS subfamily of type II transmembrane serine-proteases (TTSP) [90]. The TMPRSS13 encodes a typical TTSP structure with a transmembrane domain near the N-terminus and a trypsin-like serine protease in the extracellular domain at the C-terminal side. The role of numerous proteins is controlled by proteases, and altered expression and activity of proteases is a key event in cancer, particularly in relation to invasion, modification of the extracellular matrix and metastasis [91].

The detected mutations still need to be validated and their effects on cell growth and metastasis studied.

We have studied the pathways of these genes using the PANTHER gene ontology website: http://www.pantherdb.org/. figure 5.3 shows the different pathways of these genes.

Number	Patients	Mutation	Count	Coverage	Frequency	Control Coverage
1	66	233A>G	110	111	99.1	50
		230C>G	103	109	94.5	48
		194_208delCTCCAGGCCGGGCAT	76	117	64.96	41
2	73	248_262delAGGCATCTCCAGCCC	77	115	66.96	41
3	84	194_208delCTCCAGGCCGGGCAT	116	170	68.24	16
4	90	230C>G	87	89	97.75	25
		194_208delCTCCAGGCCGGGCAT	57	80	71.25	24
5	97A	248_262delAGGCATCTCCAGCCC	21	78	26.92	45
6	297	248A>G	22	49	44.9	54
		233A>G	17	51	33.33	54
		230C>G	15	52	28.85	52
		263G>T	14	57	24.56	54
7	364	233A>G	103	104	99.04	47
		230C>G	101	107	94.39	48
		194_208delCTCCAGGCCGGGCAT	75	106	70.75	37
8	378	233A>G	102	106	96.23	42
		230C>G	96	100	96	43
		194_208delCTCCAGGCCGGGCAT	71	111	63.96	33
9	430	248_262delAGGCATCTCCAGCCC	50	81	61.73	46

***Table 5.3**. **Types of somatic mutations detected in TMPRSS13 gene in 9 UPS patients**. The table shows the count, coverage, frequency and the control coverage of each mutation*

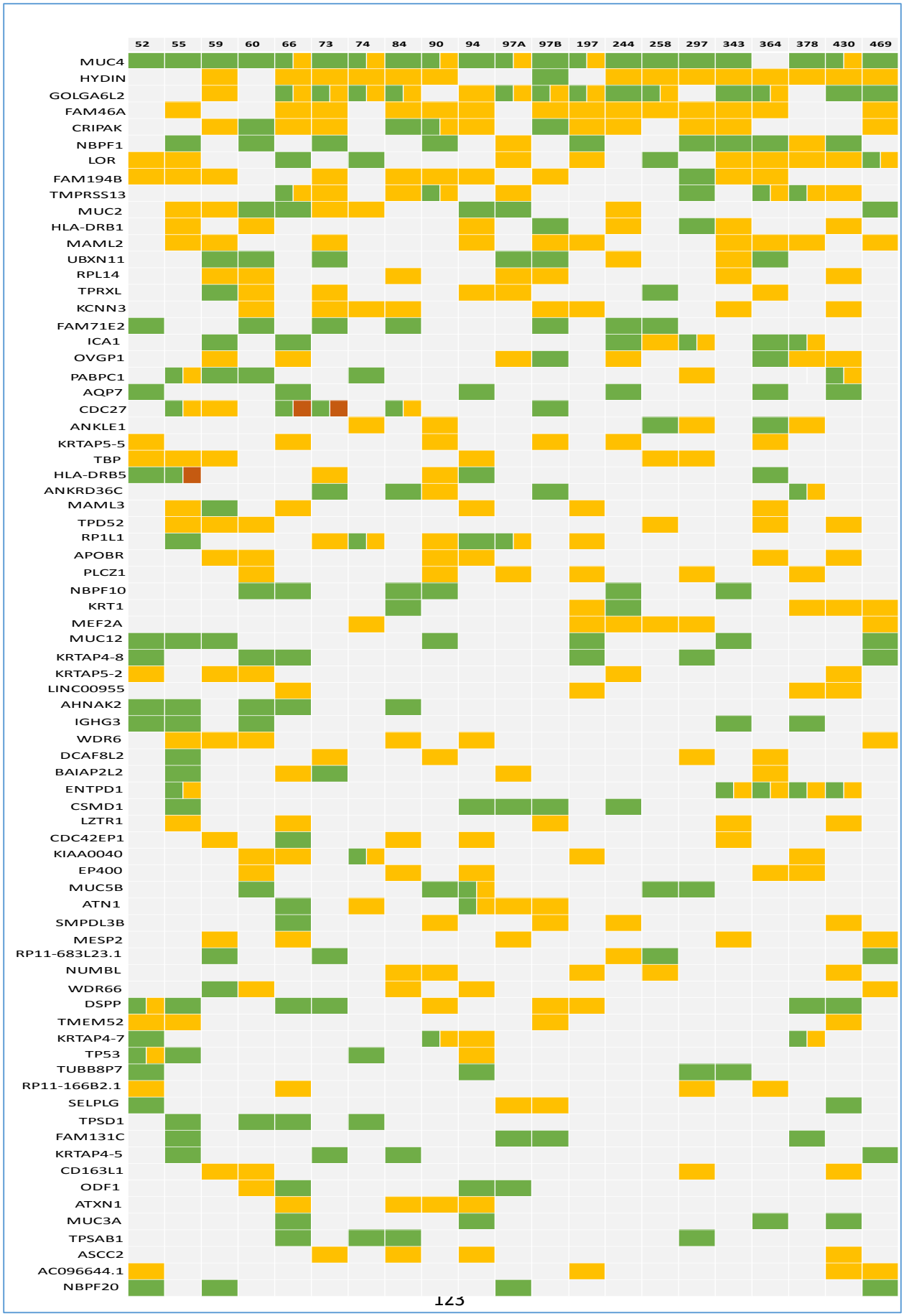

52
55
59
60
66
73
74
84
90
94
97A
97B
197
244
258
297
343
364
378
430
469
MUC4
HYDIN
GOLGA6L2
FAM46A
CRIPAK
NBPF1
LOR
FAM194B
TMPRSS13
MUC2
HLA-DRB1
MAML2
UBXN11
RPL14
TPRXL
KCNN3
FAM71E2
ICA1
OVGP1
PABPC1
AQP7
CDC27
ANKLE1
KRTAP5-5
TBP
HLA-DRB5
ANKRD36C
MAML3
TPD52
RP1L1
APOBR
PLCZ1
NBPF10
KRT1
MEF2A
MUC12
KRTAP4-8
KRTAP5-2
LINC00955
AHNAK2
IGHG3
WDR6
DCAF8L2
BAIAP2L2
ENTPD1
CSMD1
LZTR1
CDC42EP1
KIAA0040
EP400
MUC5B
ATN1
SMPDL3B
MESP2
RP11-683L23.1
NUMBL
WDR66
DSPP
TMEM52
KRTAP4-7
TP53
TUBB8P7
RP11-166B2.1
SELPLG
TPSD1
FAM131C
KRTAP4-5
CD163L1
ODF1
ATXN1
MUC3A
TPSAB1
ASCC2
AC096644.1
NBPF20

Figure 5.2. List of genes which were mutated in ⩾ 4 UPS patients with mutations ⩾ 15% frequency (count/coverage) in the tumour sequences. *Each non-synonymous SNV mutation is represented by a green square. The SNV that changing the amino acid to a stop codon is represented by a red square. The gold squares represent the Indels. When the patient has two types of mutations the square is divided into two colours according to the mutation type.*

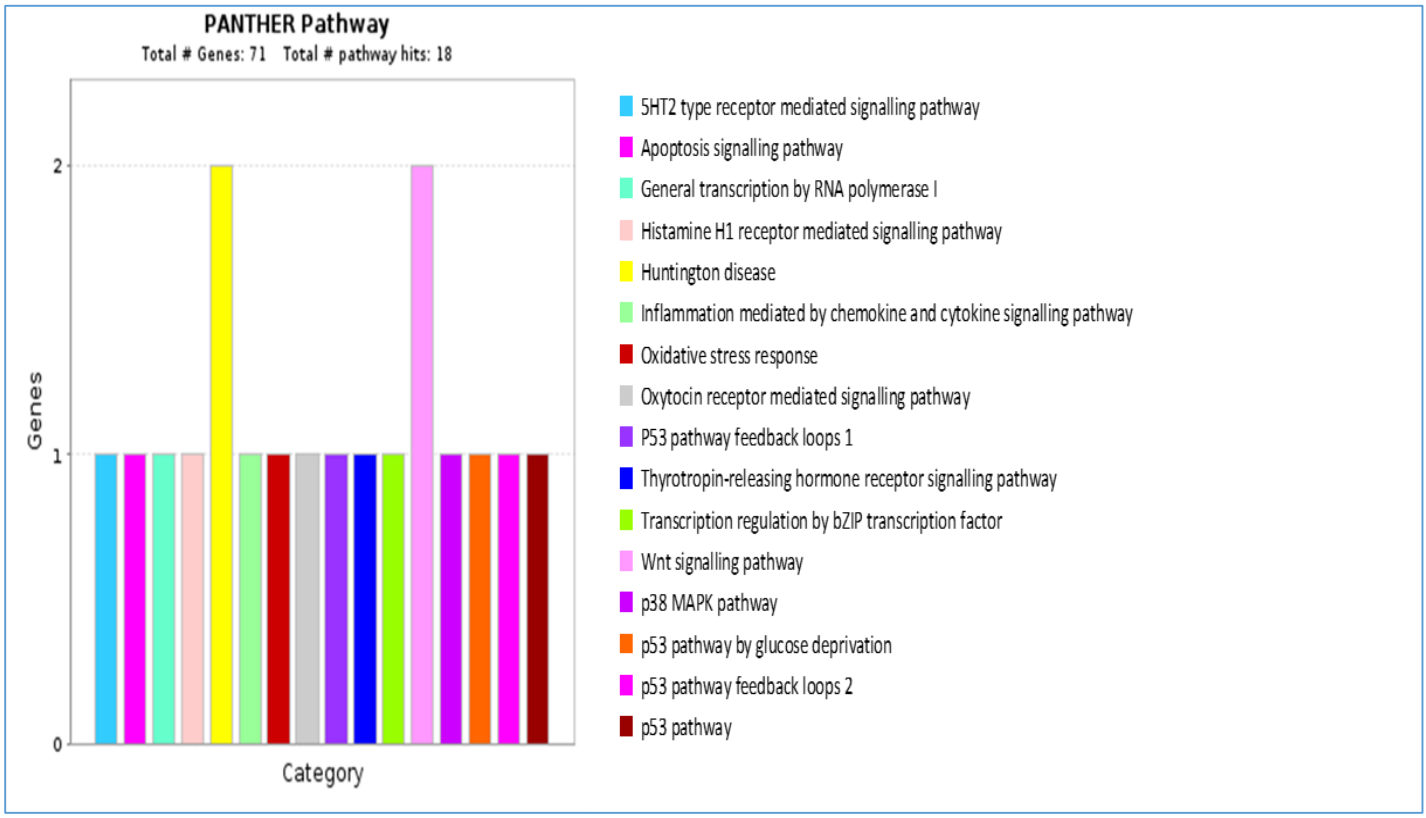


Figure 5.3. PANTHER pathways of the common mutated genes in UPS patients. *Genes which were mutated in ⩾ 4 UPS patients with mutations ⩾ 15% frequency, were listed in PANTHER gene ontology website (http://www.pantherdb.org/) to investigate their pathways.*

5.2.5 The common type of mutation in each sarcoma tumour

We have counted the number of each mutation type (mutation signature) of the SNVs in each tumour to see which is most frequent type of mutation in each tumour and if there is any similarity in mutation types between the 20 UPS patients. As shown in figure 5.4 below, all the sarcoma tumours (except tumour 74) have C>T mutation type with the highest frequency, and A>G with the second highest frequency. Tumour 74 is the only tumour that has a different mutation signature frequency, with C>A as the highest signature and C>T as the second highest.

The C>T mutation type is consistent with the deaminase activity of the apolipoprotein B mRNA editing enzyme catalytic polypeptide like (APOBEC) family. The APOBEC family is comprised of a series of molecules with conserved cytidine deaminase domains, including APOBEC1, APOBEC2, APOBEC3A to H and APOBEC4 [92]. Initially, this family of enzymes was identified as mutators of viral DNA. Most members are involved in the hypermutation state of viral genomes including retroviruses, hepatitis B virus (HBV) and human papilloma virus [93]. It was reported that the deletion of APOBEC3B (A3B) gene leads to susceptibility to HBV infection and hepatocellular carcinoma [93].

Recent studies have shown that A3B is overexpressed in multiple kinds of cancers, including breast, head and neck, lung, bladder and cervical cancers [93]. Genomic sequencing showed that C to T mutations in designated regions, regarded as a hallmark of A3B activity, is commonly found in many kinds of human cancers [93]. Knockdown of A3B showed that the cytosine deamination mutation is consistent and dependent upon A3B level in breast cancer, providing evidence that A3B is responsible for tumour mutagenesis [94].

It was demonstrated that expression of A3B was higher in chondrosarcoma cancer tissues, compared to that in normal tissues [92].

We need to test the expression of A3B in the UPS tumours and compare it to that in normal tissues.

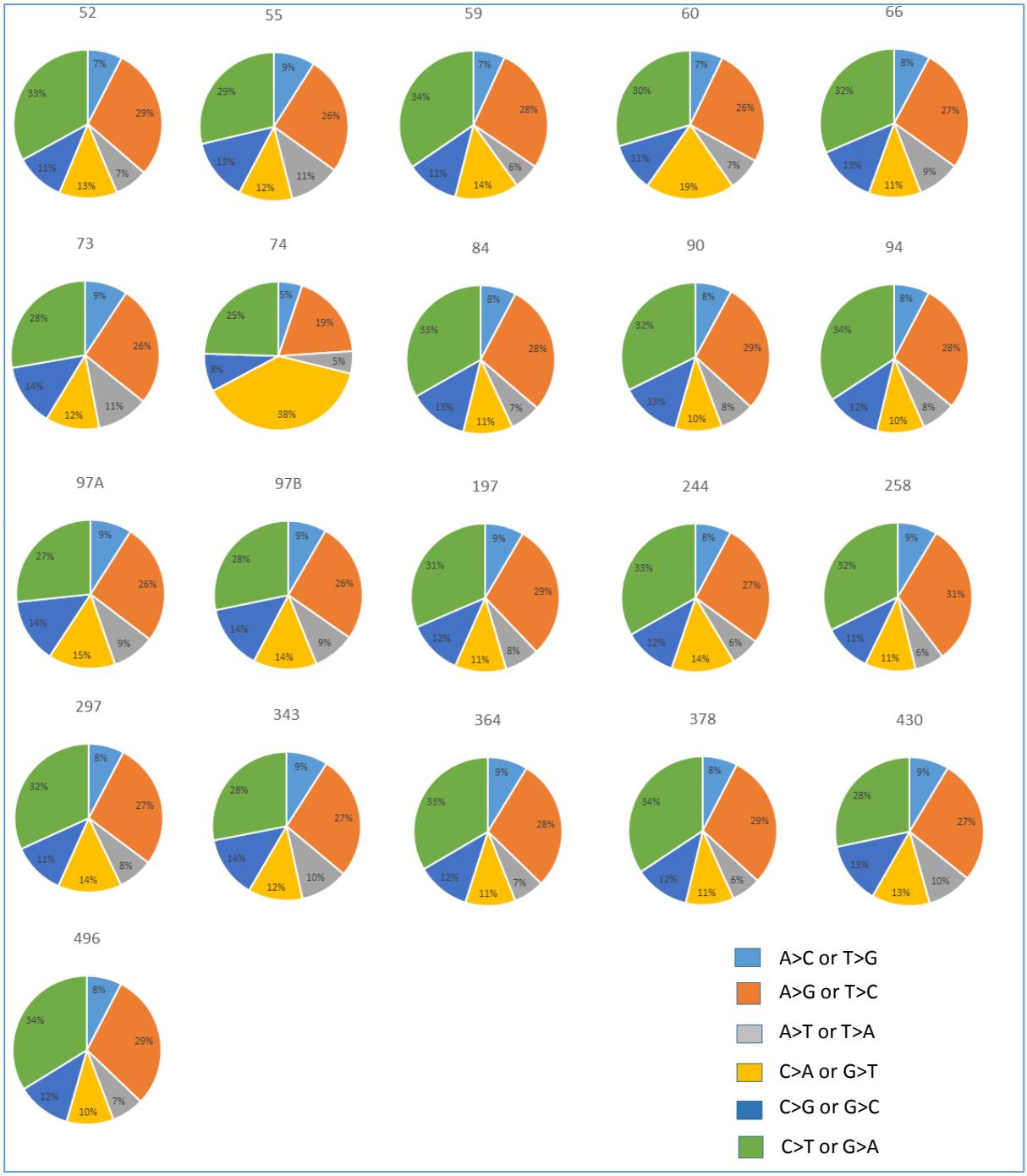


Figure 5.4. The percentages of mutation types in each UPS sample. *All of the samples (except 74) have the highest percentage of C>T mutation type represented by the green colour in the figure*.

5.2.6 Cancer heterogeneity

5.2.6.1 Mutations detected in two different regions of tumour 97

We extracted DNA from two different regions of the tumour 97 (A and B) and sent them for whole exome sequencing. We compared the two sequences with sequences of normal adjacent tissue. Tumour 97A has a smaller number of mutations detected as shown in table 5.1 above. From figure 5.2 the differences in genes detected in the two regions are obvious. The mutation signature frequency is almost the same between the two regions. When we compared the mutations detected in both regions, we found only a very small number of mutations that are common between the two regions, as shown in the Venn diagram in figure 5.5, and most of the mutations are specific to one region of the tumour; this is known as tumour heterogeneity. These two regions of the tumour may represent different grades of the tumour stage such as primary and metastatic. The common mutations (listed in table 5.4) may have arisen in the primary tumour and selection has maintained them because they may be advantageous to the tumour growth and metastasis.

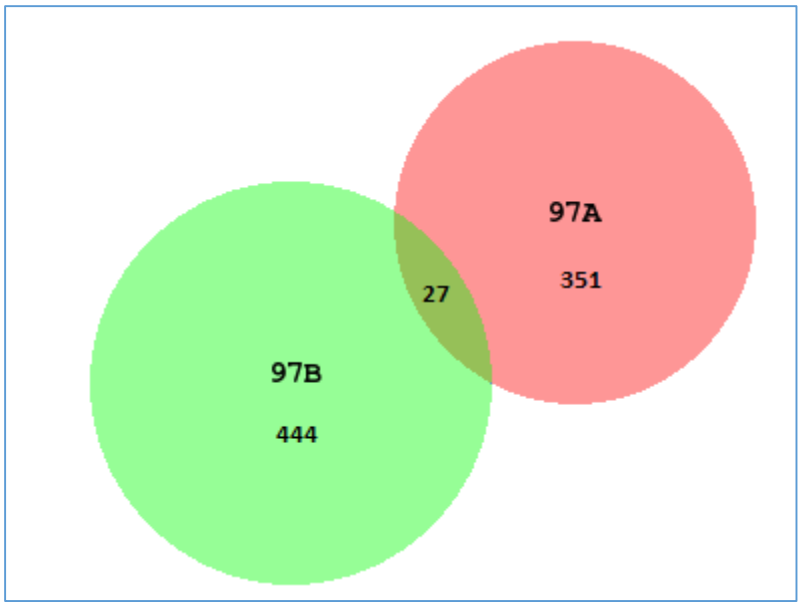


Figure 5.5. A Venn diagram of the somatic mutations detected from two regions of tumour 97 A and B.

No.	Chromosome	Region	Type	Coding region change	Count	Coverage	Frequency	Control coverage	Gene Name
1	1	62675645..62675674	Deletion	1199_1228del	104	134	77.61	32	*L1TD1*
2	3	40503520	Replacement	445delAinsGCTGCTG	26	44	59.09	23	*RPL14*
3	1	158261122	SNV	62T>C	78	165	47.27	53	*CD1C*
4	6	45390504	SNV	191C>A	16	35	45.71	23	*RUNX2*
5	12	109017651..109017680	Deletion	452_481del	40	118	33.9	55	*SELPLG*
6	11	17553007	SNV	187C>T	16	48	33.33	21	*USH1C*
7	2	75115073	SNV	2263C>T	13	47	27.66	22	*HK2*
8	5	45262297	SNV	2399T>C	40	145	27.59	91	*HCN1*
9	8	3855495	SNV	748T>G	20	78	25.64	29	*CSMD1*
10	3	10280462	SNV	1504C>T	26	103	25.24	76	*IRAK2*
11	8	97621679..97621705	Deletion	422_448del	26	137	18.98	61	*SDC2*
12	9	90501486	SNV	2084G>C	25	133	18.8	46	*SPATA31E1*
13	3	65342333	SNV	418G>A	24	131	18.32	72	*MAGI1*
14	9	19116634	SNV	926G>A	18	106	16.98	30	*PLIN2*
15	10	50535007^50535008	Insertion	2106_2107insACACACAC	8	51	15.69	16	*C10orf71*
16	1	16388642	SNV	220C>T	5	32	15.62	8	*FAM131C*
17	16	1478488..1478561	Deletion	145-55_163del	14	104	13.46	11	*C16orf91*
18	7	139281691	SNV	2489G>A	16	165	9.7	39	*HIPK2*
19	7	142224264	SNV	4A>G	3	37	8.11	24	*TRBV11-1*
20	7	119915308	SNV	622C>G	25	309	8.09	78	*KCND2*
21	3	195511918	SNV	6533C>T	2	25	8	8	*MUC4*
22	6	10756712	SNV	293+1247C>T	6	86	6.98	33	*TMEM14B , RP11-637O19.3 , SYCP2L*
23	6	10756728	SNV	322C>T	6	87	6.9	30	*TMEM14B , RP11-637O19.3 , SYCP2L*
24	15	41862354..41862436	Deletion	1247_1248-2del	17	252	6.75	17	*TYRO3*
25	12	7045894	SNV	1464G>T	5	78	6.41	38	*ATN1*

26	12	53045645..53045646	Replacement	281_282delGTinsC	2	37	5.41	18	*KRT2*
27	7	142224256	SNV	12G>C	2	40	5	27	*TRBV11-1*

Table 5.4. The common somatic mutations in regions A and B of tumour 97

5.2.6.2 Pathways of the mutated genes in the two regions of tumour 97

We have investigated the pathways of the genes with somatic mutations in regions A and B of the tumour 97 by using PANTHER gene ontology website: (http://www.pantherdb.org/). The pathways of mutated genes detected in region A and region B are shown in figure 5.6 and 5.7 respectively. Although most of the mutated genes are different between the two regions, the two regions share many pathways of their mutated genes, as shown in figures 5.6 and 5.7. The two highest pathways are the same in the two regions: Cadherin signalling pathway and Wnt signalling pathway. There are other common pathways such as: T cell activation, Ras Pathway, FGF signalling pathway, Inflammation mediated by chemokine and cytokine signalling pathway and Ubiquitin proteasome pathway.

There are some pathways that are specific to each region; for example: p53 pathway and B cell activation in region A, and Apoptosis signalling pathway and Cytoskeletal regulation by Rho GTPase for region B.

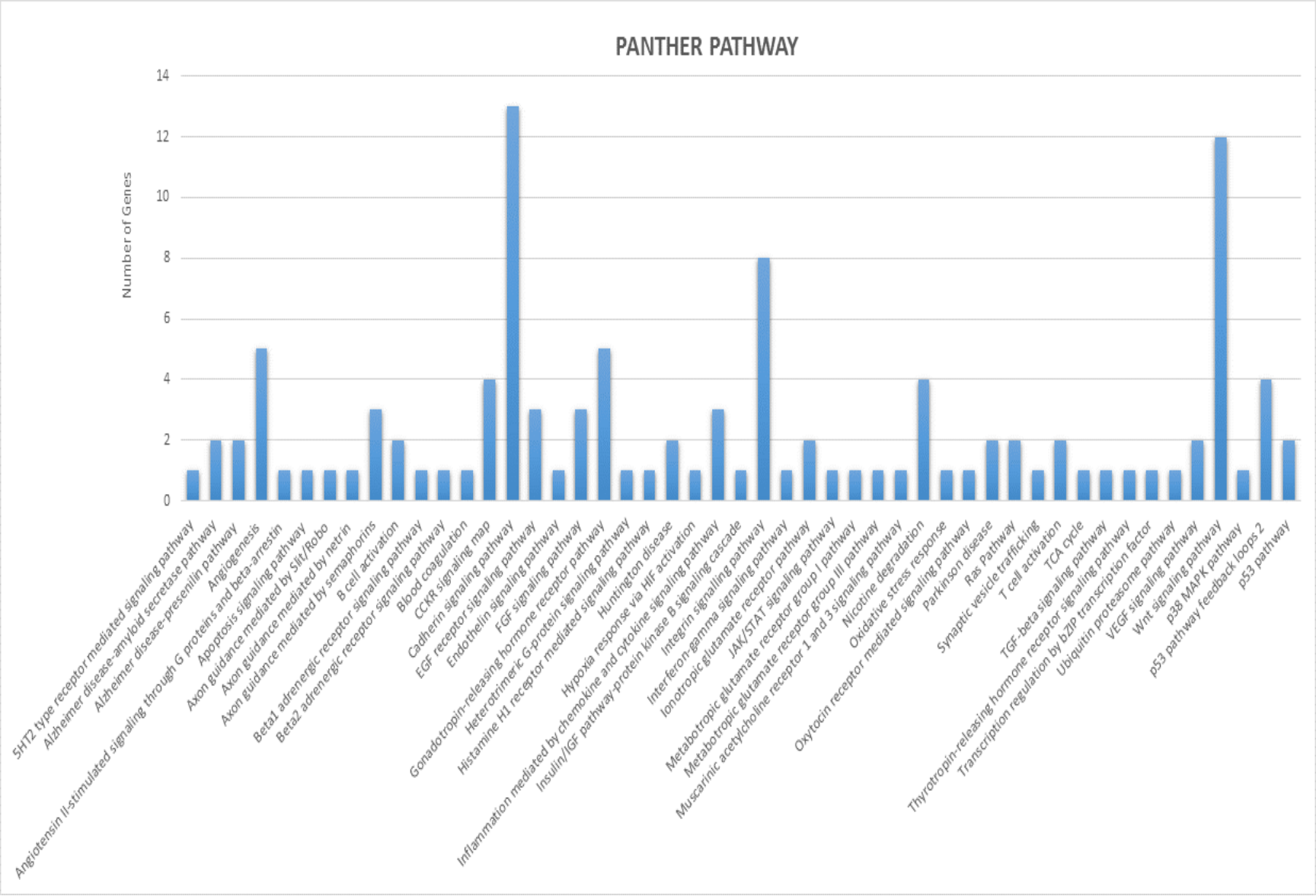

PANTHER PATHWAY
Number of Genes
0
2
4
6
8
10
12
14
5HT2 type receptor mediated signaling pathway
Alzheimer disease-amyloid secretase pathway
Alzheimer disease-presenilin pathway
Angiogenesis
Angiotensin II-stimulated signaling through G proteins and beta-arrestin
Apoptosis signaling pathway
Axon guidance mediated by Slit/Robo
Axon guidance mediated by netrin
Axon guidance mediated by semaphorins
B cell activation
Beta1 adrenergic receptor signaling pathway
Beta2 adrenergic receptor signaling pathway
Blood coagulation
CCKR signaling map
Cadherin signaling pathway
EGF receptor signaling pathway
Endothelin signaling pathway
FGF signaling pathway
Gonadotropin-releasing hormone receptor pathway
Heterotrimeric G-protein signaling pathway
Histamine H1 receptor mediated signaling pathway
Huntington disease
Hypoxia response via HIF activation
Inflammation mediated by chemokine and cytokine signaling pathway
Insulin/IGF pathway-protein kinase B signaling cascade
Integrin signalling pathway
Interferon-gamma signaling pathway
Ionotropic glutamate receptor pathway
JAK/STAT signaling pathway
Metabotropic glutamate receptor group I pathway
Metabotropic glutamate receptor group III pathway
Muscarinic acetylcholine receptor 1 and 3 signaling pathway
Nicotine degradation
Oxidative stress response
Oxytocin receptor mediated signaling pathway
Parkinson disease
Ras Pathway
Synaptic vesicle trafficking
T cell activation
TCA cycle
TGF-beta signaling pathway
Thyrotropin-releasing hormone receptor signaling pathway
Transcription regulation by bZIP transcription factor
Ubiquitin proteasome pathway
VEGF signaling pathway
Wnt signaling pathway
p38 MAPK pathway
p53 pathway feedback loops 2
p53 pathway

Figure 5.6. PANTHER pathways of the list of genes with somatic mutations in the tumour 97A*. The PANTHER gene ontology website (http://www.pantherdb.org/) was used to investigate the pathways of the genes with somatic mutations detected when the sequence of the region A of the tumour 97 compared to the sequence of the normal adjacent tissue.*

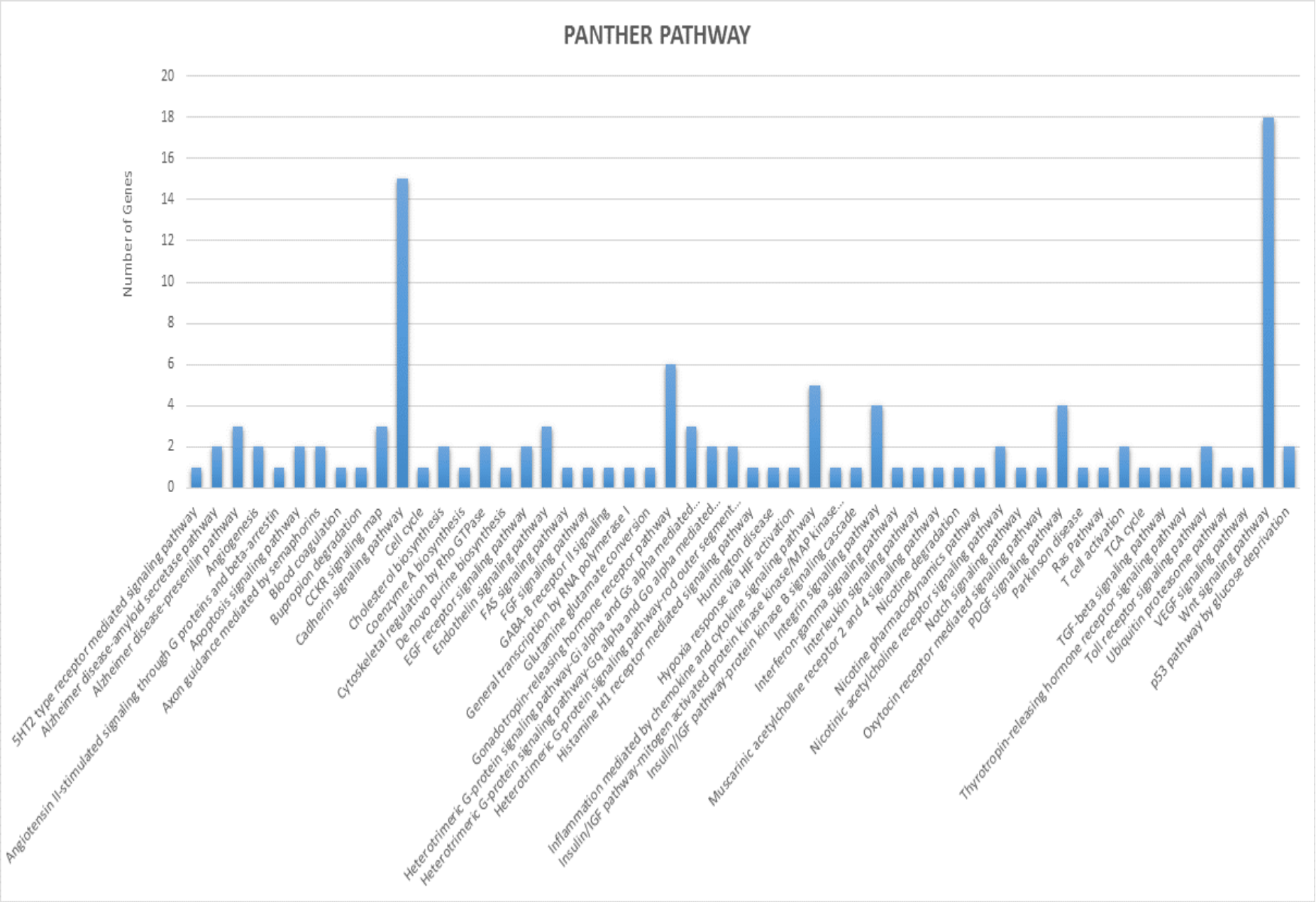

PANTHER PATHWAY
Number of Genes
0
2
4
6
8
10
12
14
16
18
20
5HT2 type receptor mediated signaling pathway
Alzheimer disease-amyloid secretase pathway
Alzheimer disease-presenilin pathway
Angiogenesis
Angiotensin II-stimulated signaling through G proteins and beta-arrestin
Apoptosis signaling pathway
Axon guidance mediated by semaphorins
Blood coagulation
Bupropion degradation
CCKR signaling map
Cadherin signaling pathway
Cell cycle
Cholesterol biosynthesis
Coenzyme A biosynthesis
Cytoskeletal regulation by Rho GTPase
De novo purine biosynthesis
EGF receptor signaling pathway
Endothelin signaling pathway
FAS signaling pathway
FGF signaling pathway
GABA-B receptor II signaling
General transcription by RNA polymerase I
Glutamine glutamate conversion
Gonadotropin-releasing hormone receptor pathway
Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated...
Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated...
Heterotrimeric G-protein signaling pathway-rod outer segment...
Histamine H1 receptor mediated signaling pathway
Huntington disease
Hypoxia response via HIF activation
Inflammation mediated by chemokine and cytokine signaling pathway
Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase...
Insulin/IGF pathway-protein kinase B signaling cascade
Integrin signalling pathway
Interferon-gamma signaling pathway
Interleukin signaling pathway
Muscarinic acetylcholine receptor 2 and 4 signaling pathway
Nicotine degradation
Nicotine pharmacodynamics pathway
Nicotinic acetylcholine receptor signaling pathway
Notch signaling pathway
Oxytocin receptor mediated signaling pathway
PDGF signaling pathway
Parkinson disease
Ras Pathway
T cell activation
TCA cycle
TGF-beta signaling pathway
Thyrotropin-releasing hormone receptor signaling pathway
Toll receptor signaling pathway
Ubiquitin proteasome pathway
VEGF signaling pathway
Wnt signaling pathway
p53 pathway by glucose deprivation

Figure 5.7. PANTHER pathways of the list of genes with somatic mutations in the tumour 97B*. The PANTHER gene ontology website (http://www.pantherdb.org/) was used to investigate the pathways of the genes with somatic mutations detected when the sequence of the region B of the tumour 97 compared to the sequence of the normal adjacent tissue.*

Tumour heterogeneity refers to the existence of subpopulations of cells with distinct genotypes and phenotypes, that may harbour divergent biological behaviours, within a primary tumour and its metastases, or between tumours of the same histopathological subtype (intra- and inter-tumour, respectively) [95].

Cancers of all types are now recognised to consist of highly diverse populations of cells, where intra-tumour heterogeneity is detectable at the genetic, epigenetic, and phenotypic levels [96]. A study has assessed the mutational spectrum in a patient with metastatic lobular breast cancer and identified that out of 32 coding mutations, 19 were specific to metastasis [23]. In this context, it was shown that performing single biopsies of primary tumours or metastatic deposits is unlikely to reveal the complete profile of genomic alterations in any tumour [97]. This multiregional separation of molecular aberrations can lead to sampling bias, potentially impairing the interpretation of the molecular characterization of tumours and having an impact on the selection of treatment [97]. For drug target development, clonal heterogeneity must be taken into account, as subclones can compete and synergize for growth in a symbiotic manner [98].

5.2.7 Comparison of the somatic mutations detected by CLC and MuTect

We have compared the variants detected by the CLC analysis and the variants detected by another point mutation caller; MuTect (MuTect analysis was performed by Graeme Grimes, MRC Human Genetics Unit). MuTect is a method developed at the Broad Institute, for the reliable and accurate identification of somatic point mutations in NGS data of cancer genomes. MuTect takes as input sequence data from matched tumour and normal DNA, after alignment of reads to a reference genome and standard pre-processing steps that include marking of duplicate reads, recalibration of base quality scores and local realignment. The method operates on each genomic locus independently and consists of four key steps (Fig 5.8): (i) Removal of low-quality sequence data; (ii) variant detection in the tumour using a Bayesian classifier; (iii) filtering to remove false positives resulting from correlated sequencing artefacts that are not captured by the error model; and (iv) designation of the variants as somatic or germline by a second Bayesian classifier [99].

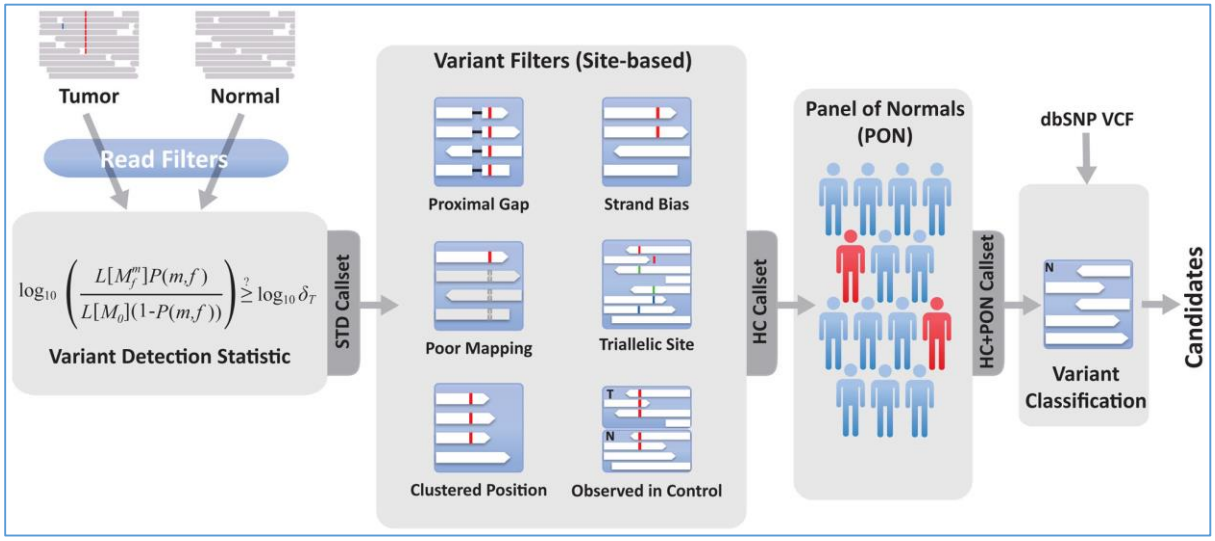


Figure 5.8. Overview of the detection of a somatic point mutation using MuTect. *MuTect takes as input NGS data from tumour and normal samples and, after removing low-quality reads, determines whether there is evidence for a variant beyond the expected random sequencing errors. Candidate variant sites are then passed through six filters to remove artefacts. Next, a panel of normal samples filter is used to screen out remaining false positives caused by rare error modes only detectable in additional samples. Finally, the somatic or germline status of passing variants is determined using the matched normal sample [99].*

The CLC Workbench detects SNVs and Indels, but MuTect detects only SNVs, which makes CLC advantageous over MuTect. The number of SNVs detected by CLC is much greater than the number of SNVs detected by MuTect in each patient as shown in table 5.5 below. This is may be due to the fact that the CLC detects variants at very low frequency, which MuTect cannot detect.

No.	Patient	No. of SNVs detected by CLC	No. of variants detected by MuTect	No. of variants detected by both callers
1	52	412	305	126
2	55	399	169	148
3	59	242	48	25
4	60	359	67	41
5	66	293	56	31
6	73	274	47	30
7	74	506	148	49
8	84	364	58	24
9	90	312	140	80
10	94	434	215	182
11	97A	281	21	12
12	97B	346	22	20
13	197	207	44	20
14	244	237	48	37
15	258	195	62	32
16	297	297	60	44
17	343	311	55	43
18	364	233	59	32
19	378	239	79	29
20	430	423	31	22
21	496	152	21	7

Table 5.5. The number of somatic mutations detected by the CLC software and MuTect in UPS patients.

The number of shared variants detected by both methods is low in each patient as shown in table 5.5 above, which means that each caller method detects a high number of variants which are not detected by the other method.

Mutations detected in *Tp53* gene by the two methods are examples, as shown in table 5.6 below. CLC detected five mutations in *Tp53*, a deletion of 15 bases and a 1-base deletion in patients 52 and 94 respectively, leading to frameshift mutations; and three SNVs in patients 52, 55 and 74. The deletions were not detected by MuTect because it does not detect Indels, as mentioned above. The three SNVs detected by CLC were also detected by MuTect. In contrast, MuTect detected another SNV in patient 90 with high frequency and high coverage in the tumour and normal samples, but this SNV was not detected by the CLC.

When we looked at this mutation site in the CLC genome browser of patient 90, we could see that the mutation was there, as shown in figure 5.9, but the count is much less than the count that was detected by MuTect, which is only 4 out of 82 reads in the tumour which makes the frequency less than 5%; that is why this mutation was not called out by the CLC.

Mutations detected in Tp53 by CLC method:									
Patient	**Region**	**Type**	**Reference**	**Allele**	**Count**	**Coverage**	**Frequency %**	**Control coverage**	**Amino acid change**
52	7577596..7577610	Deletion	AGTCAGAGCCAACCT	-	34	82	41.46	22	Gly226fs
52	7577536	SNV	T	C	22	81	27.16	23	Arg249Gly
55	7577538	SNV	C	T	72	75	96	23	Arg248Gln
74	7578265	SNV	A	G	42	136	30.88	82	Ile195Thr
94	7579508	Deletion	G	-	40	82	48.78	22	Pro60fs
Mutations detected in Tp53 by MuTect method:									
52	7577536	SNV	T	C	49	70	70	20	Arg249Gly
55	7577538	SNV	C	T	66	69	95.6	21	Arg248Gln
74	7578265	SNV	A	G	84	127	66.1	78	Ile195Thr
90	7577121	SNV	G	A	117	127	92	35	Arg273Cys

Table 5.6. Somatic mutations detected in TP53 gene by the CLC software and MuTect in UPS patients.

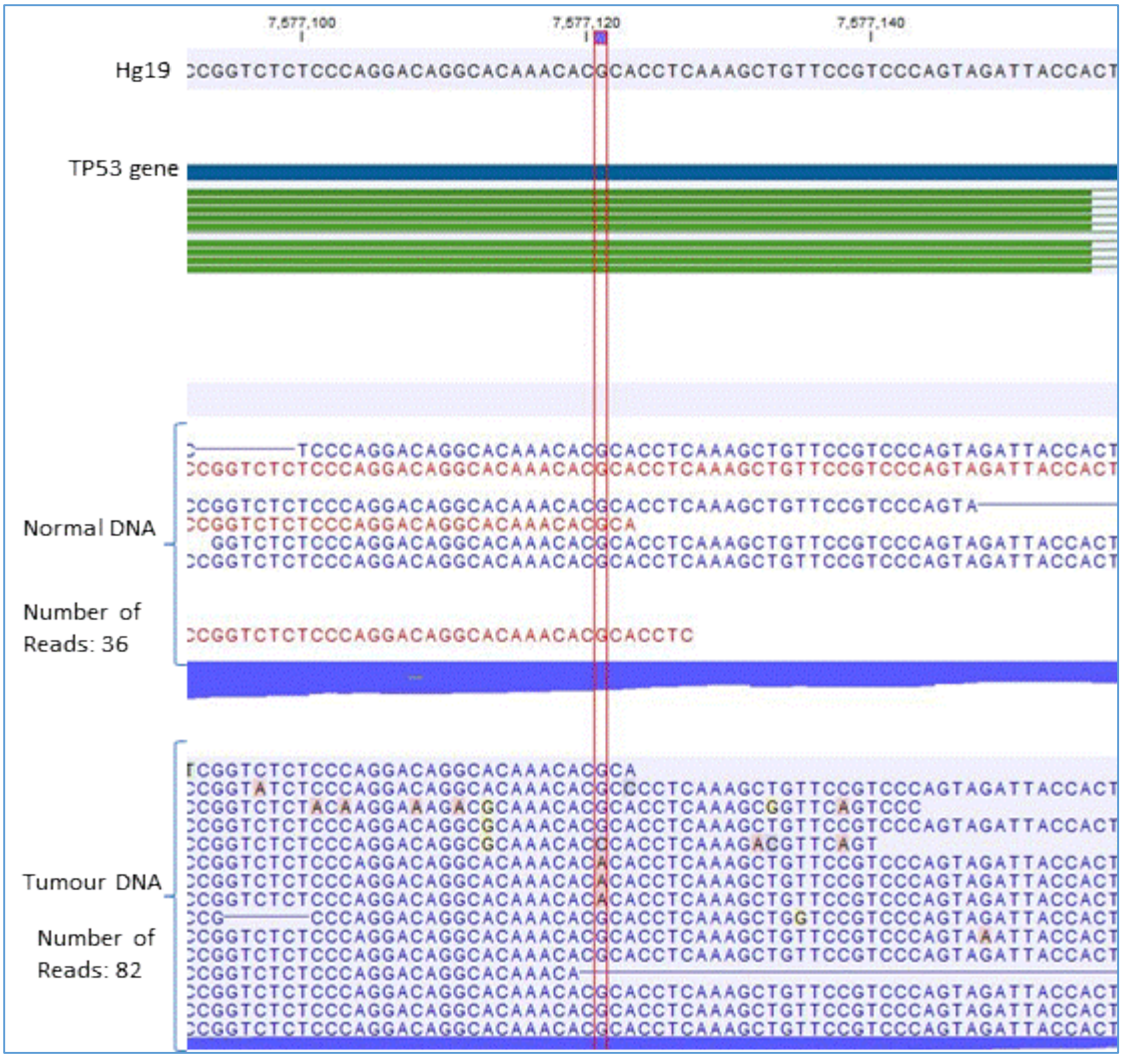


Figure 5.9. The CLC genomic browser of the G>A mutation in TP53 gene in patient 90*. The mutation presents in very low number in the tumour sequences (4/82), making its frequency lower than 5%.*

In patient 55, The CLC detected 148 variants out of the 169 variants detected by MuTect, as shown in the Venn diagram in figure 5.10. In other words, 87.5% of the somatic variants detected by MuTect have been detected by the CLC. Table 5.7 lists the rest of the variants (21 variants) that have not been detected by the CLC, with the reasons for their not being detected. Nine of them have the variant in the normal reads, which is why the CLC did not detect them.

There are many SNVs detected by the CLC (251) but not detected by MuTect. Most of these variants have high frequency and high coverage in the tumour and normal samples, as in the

mutation in *KIAA1958* gene, shown in figure 5.11, which has a frequency of 87.6% and coverage of 234 and 100 in tumour and normal samples respectively.

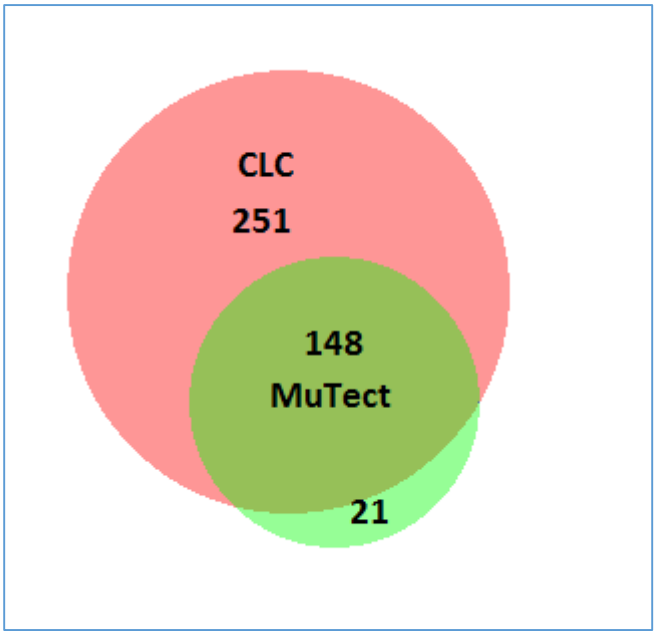


Figure 5.10. A Venn diagram showing the number of mutations detected by the CLC and MuTect in patient 55, and the number of mutations common to both methods.

No.	Gene name	Chromosome	Start Position	Reference Allele	Tumour Allele	Mutation	Amino Acid change	Reason for not detected in CLC
1	*AMY1C*	1	104297162	C	T	c.920C>T	Ala307Val	False negative
2	*RETSAT*	2	85570849	C	T	c.1606G>A	Gly536Arg	Detected in normal
3	*RETSAT*	2	85570857	G	A	c.1598C>T	Ala533Val	Detected in normal
4	*POTEJ*	2	131414975	A	T	c.2642A>T	Tyr881Phe	Detected in normal
5	*MAATS1*	3	119463011	C	T	c.1870C>T	Arg624Cys	False negative
6	*TACC3*	4	1732978	G	A	c.1541G>A	Gly514Glu	Low frequency
7	*PCLO*	7	82784471	A	G	c.1486T>C	Ser496Pro	Detected in normal
8	*OR2AE1*	7	99473692	C	T	c.965G>A	Arg322Gln	Low frequency
9	*PRSS1*	7	142459626	C	T	c.202C>T	Arg68Cys	Detected in normal
10	*FAM115C*	7	143400205	G	A	c.118G>A	Val40Met	Detected in normal
11	*PABPC1*	8	101724606	G	A	c.956C>T	Thr319Ile	Detected in normal
12	*GLIS3*	9	3828360	C	T	c.2705G>A	Arg902His	False negative
13	*FAM154A*	9	18928669	C	T	c.806G>A	Cys269Tyr	Low frequency
14	*IFNA14*	9	21239468	T	C	c.467A>G	Glu156Gly	No variant detected
15	*IFNA14*	9	21239504	T	C	c.431A>G	Lys144Arg	False negative
16	*IFNA14*	9	21239589	T	G	c.346A>C	Met116Leu	Detected in normal
17	*OR8G5*	11	124135536	G	A	c.814G>A	Ala272Thr	Detected in normal
18	*GOLGA8T*	15	30433598	A	T	c.844A>T	Arg282Trp	False negative
19	*ODF3L2*	19	464067	C	T	c.647G>A	Arg216His	False negative
20	*EDEM2*	20	33722572	C	T	c.671G>A	Arg224His	False negative
21	*FAM83C*	20	33880040	A	C	c.68T>G	Val23Gly	Bad sequence

Table 5.7. Variants detected by MuTect only; 21 variants were detected by MuTect and not by CLC. *Includes their details and possible reasons for not being detected by the CLC.*

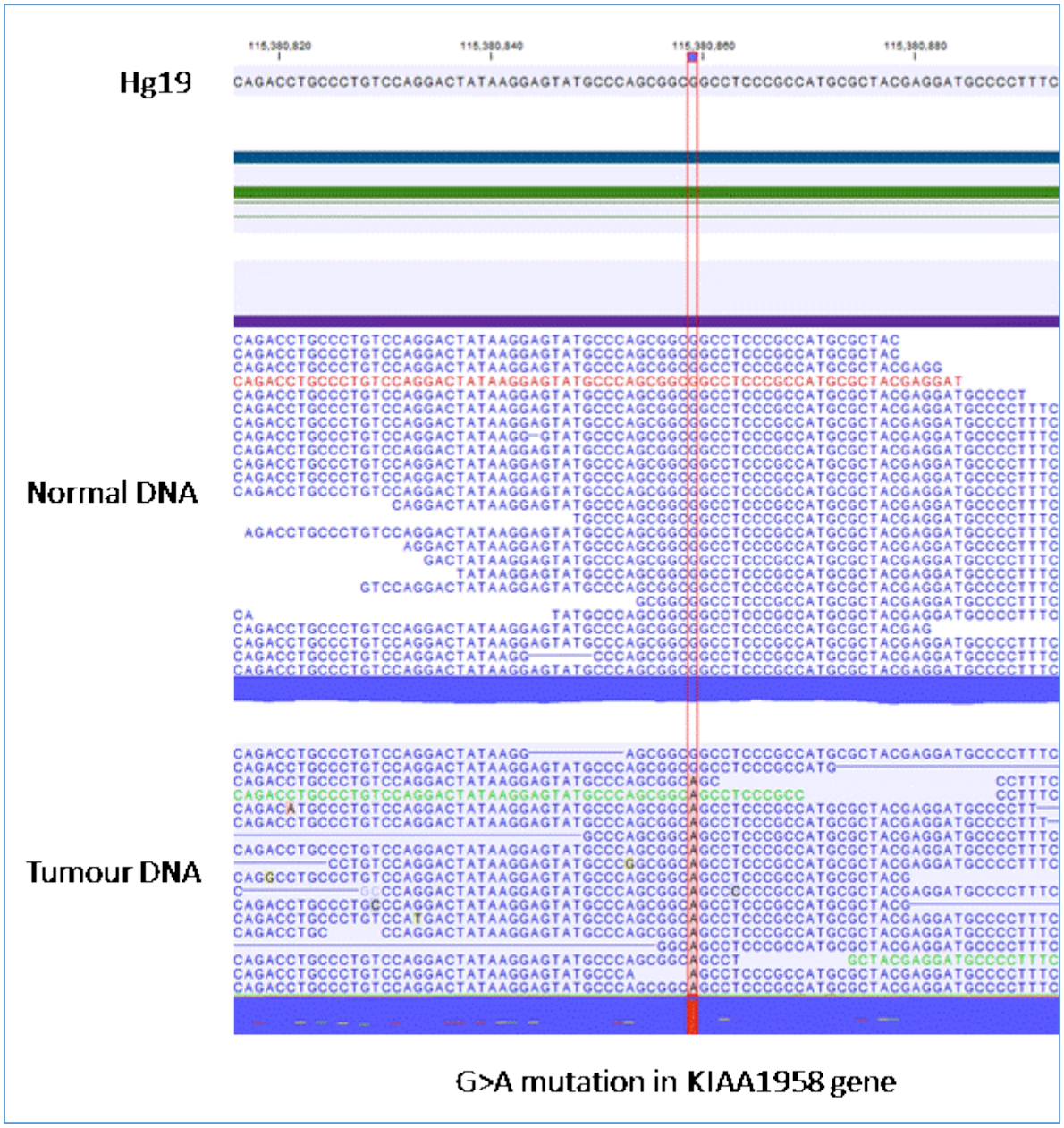


***Figure 5.11*. CLC genome browser of a G>A mutation in KIAA1958**. *The mutation presents with high frequency in the tumour reads, and there is no mutation in the normal reads.*

A previous study has used whole genome sequencing of a melanoma sample and matched blood, whole exome sequencing of 18 lung tumour–normal pairs and seven lung cancer cell lines to evaluate MuTect and other mutation caller methods such as VarScan 2 [100]. Out of 119 true positive SNVs in melanoma, MuTect detected 115 SNVs and missed four SNVs for reasons of nearby gap events and an alternate allele in normal tissue.

For the 43 WES samples, 118 putative SNVs were validated as true positives. Of these, 92 SNVs had high coverage in the tumour and normal samples (high quality), while 26 of them had less than 8 coverage in the normal sample (low quality). MuTect identified 77 SNVs of the 92 high quality SNVs, missing 15 variants. All the other methods in the have detected more high quality variants than did MuTect. Of the 26 low quality variants, MuTect detected the largest number: 19 out of the 26 SNVs (missing 7 variants) compared to other methods, which makes MuTect the best for detection of low quality variants.

From the analysis of the 18 lung tumours, MuTect detected a total of 11 false positive SNVs, the highest number among the five tools used in the study.

5.2.8 Copy number variation (CNV) in UPS

As described in the head and neck chapter, the same tools in the CLC Biomedical Genomics Workbench have been used to detect the CNVs in the UPS samples. Many CNVs were detected, of both amplifications and deletions, in each sample (data in additional file). The tumour of patient 55, for example, has 52 CNVs detected, 19 of them are amplifications and 33 are deletions. The amplifications were in chromosomes 3, 7, 15, 19 and 21, as shown in figure 5.12 below. The 33 deletions were distributed through different chromosomes and they are in small regions, except for the two large deletions in chromosome 10.

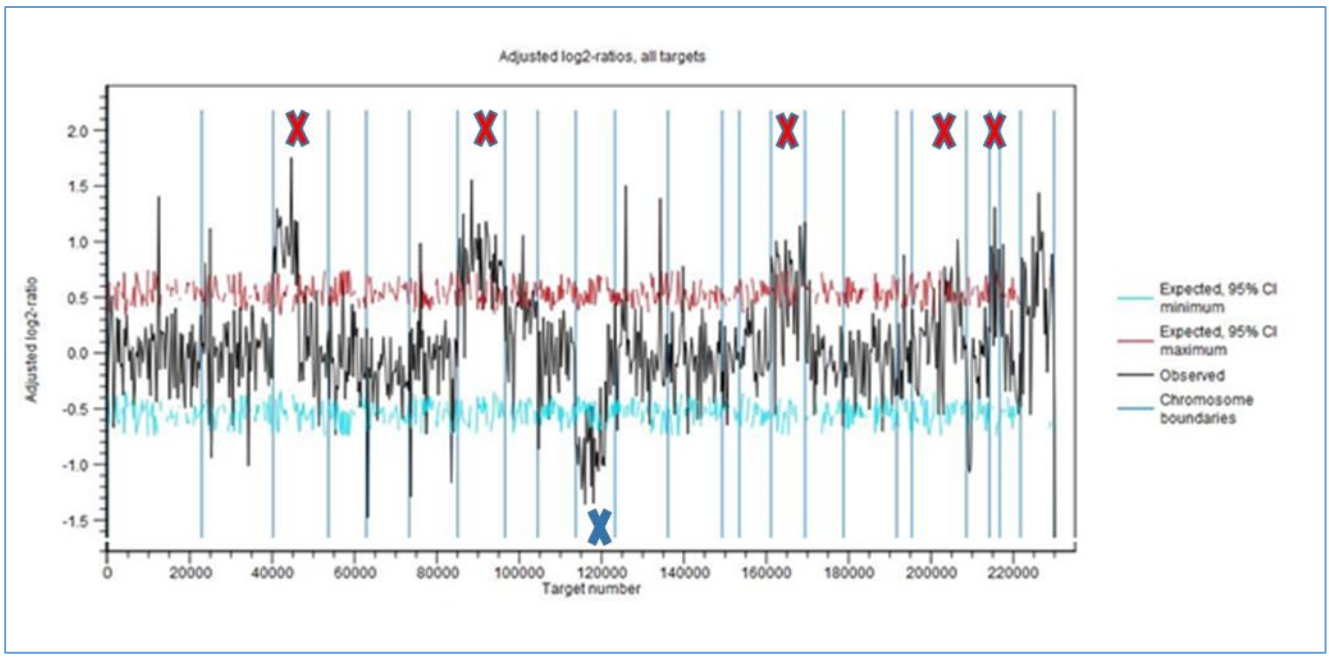


Figure 5.12. A graph showing the mean adjusted log-ratios of coverages in the report produced by the CNV detection in patient 55*. In this report, regions of the 3rd, 7th, 15th, 19th, and 21st chromosomes are amplified (red X), and regions of chromosome 10 are deleted (blue X). The log-ratios of coverages of targets on these chromosomes are significantly higher or lower than for targets on other chromosomes. The black line in these regions is outside the boundaries defined by the cyan and red lines.*

5.2.8.1 Shared CNVs in UPS samples

There are regions of 52 CNVs that are shared between two or three samples of UPS, as shown in table 4.8. Of these, 11 CNVs are deletions and 41 are amplifications. There are three CNVs detected in three UPS samples, two gains and one loss. The first gain is on chromosome 3, at location 3q13.33, detected in patients 90, 94 and 97A, with the minimum CNV length of 145928 bases; the other gain is on chromosome 5 (5q11.2) detected in patients 84, 94 and 197, with minimum length of 209501. The loss is on chromosome 11 (11p15.4) and found in patients 73, 94 and 197, with a minimum length of 944. All the other shared CNVs are only seen in two samples. Patients 90 and 97A share the highest number of CNVs with 16. Patients 84 and 94 share the second highest number, with 14 CNVs, and finally patients 94 and 97A have 9 CNVs. A study was carried out by [80] to detect CNVs in 20 UPS untreated patients using array-based comparative genome hybridization. They reported that the most frequently observed significant alterations involved gains at: 20q13.33 in 75% of samples; 1q21.3-q23.1 in 60%; 7q22.1 in 60%; 9q34.11 and 20p11.21 in 45%; and 1q21.1-1q21.2, 8p11.21, 11q13.1 and 16p13.3 in 40% of samples. They have validated the amplifications of some of the genes located in some of these regions by quantitative real-time PCR. Most of these CNVs were also detected in our analysis. There are gains detected at: 20q13.13 in patients 94 and 97A; 1q22 in patients 90 and 97A; 7q22.1 in patients 90 and 97A; 11q13.1 in patients 90 and 97A; and 16p13.3 in patients 84 and 94.There is a large difference in the number of CNVs detected in samples 97A and 97B. There are 431 CNVs detected in 97A and only 90 CNVs in 97B, and they share only four CNVs as shown in table 5.8. This means that the number and type of CNVs are also different between different sites of the same tumour.

In conclusion; the newly available software was able to identify genomic mutations in UPS and gave rise to an overlapping variant calling rate with the standard MuTect software.

Number	Samples	Chromosome location	Region	Minimum CNV length	Consequences
1	60, 84	4p16.1	7043045..10509695	3466651	Gain
2	84, 94	2p22.3	33824206..36583843	2759638	Gain
3	84, 94, 97A	5q11.2	53606160..53815660	209501	Gain
4	84, 94	5q31.2	137771222..137804146	32925	Gain
5	84, 94	7p15.2	26903889..27497459	593571	Gain
6	84, 94	8q22.1	96281271..97172921	891651	Gain
7	84, 94	10p11.21	35896547..35930313	33767	Gain
8	84, 94	10q25.2	112837732..112838496	765	Gain
9	84, 94	13q14.3	53419925..53422629	2705	Gain
10	84, 94	14q13.2	35871605..36005206	133602	Gain
11	84, 94	16p13.3	1756292..1877945	121654	Gain
12	90, 97A	1q22	155531872..155533272	1401	Gain
13	90, 94	3q13.2	111697898..111698138	241	Gain
14	90, 94, 97A	3q13.33	120169505..120315432	145928	Gain
15	90, 97A	4p15.2	26321381..26322455	1075	Gain
16	90, 97A	4q23	100867620..100871448	3829	Gain
17	90, 97A	5q11.2	54603174..54604070	897	Gain
18	90, 97A	6p21.33	31783494..31803251	19758	Gain
19	90, 97A	6p21.33	31939130..31940828	1699	Gain
20	90, 97A	7q22.1	100859979..100861570	1592	Gain
21	90, 97A	8q23.1	109260724..109455990	195267	Gain
22	90, 97A	10q23.31	91087672..91087912	241	Gain
23	90, 97A	10q24.32	104180746..104181231	486	Gain
24	90, 97A	11q13.1	63919711..63953790	34080	Gain
25	90, 97A	12p13.2	10765962..10766206	245	Gain
26	90, 97A	12q14.3	66562928..66563792	865	Gain
27	90, 94	13q14.13	45914264..45915312	1049	Gain
28	90, 94	16q21	66583885..66586779	2895	Gain
29	90, 97A	17p13.3	1303230..1303474	245	Gain
30	90, 94	17q23.3	62500715..62503327	2613	Gain
31	90, 97A	22q11.21	19466543..19467764	1222	Gain
32	94, 97A	1p36.33	861267..1565994	704728	Gain
33	94, 97A	1q31.2	193070213..193091479	21267	Gain
34	94, 97A	4q28.1	128703073..128703863	791	Gain
35	94, 97A	7q21.13	90747375..90896159	148785	Gain
36	94, 97A	8q24.13	125384090..125487563	103474	Gain
37	94, 97A	9p13.2	37784632..37800695	16064	Gain
38	94, 97A	10q23.33	94333589..94333829	241	Gain
39	94, 97A	10q24.32	104262278..104264144	1867	Gain
40	94, 97A	20q13.13	48769952..48808616	38665	Gain
41	97A, 97B	17p13.1	8192047..16252058	8060012	Gain
42	55, 364	4q13.2	69416379..69433917	17539	Loss
43	66, 84	18p11.21	12095417..12127862	32446	Loss
44	73, 94, 197	11p15.4	4976026..4976969	944	Loss
45	84, 94	4q24	107114760..107183372	68613	Loss
46	84, 94	16q12.2	55559422..55585020	25599	Loss
47	84, 94	17p11.2	16455128..16676861	221734	Loss
48	84, 94	17q12	33255059..33285772	30714	Loss
49	97A, 97B	1p36.33	721382..792446	71065	Loss
50	97A, 97B	1p13.2	112991586..113269425	277840	Loss

51	97A, 97B	Xp22.13	19008966..19037888	28923	Loss
52	244, 378	10p12.33	17756495..17891625	135131	Loss

Table 5.8. Shared CNVs in some of the UPS samples.

CHAPTER SIX
Identifying expressed somatic mutations and RNA editing events in the whole transcriptome sequence of three UPS patients

6.1 Introduction

6.1.1 The principle of RNA sequencing

The major limitation of human cancer genome sequencing is that the information does not reflect the expressed state of the cancer at the time of surgery. For example, we could consider the cancer genome as a “fossil record” that reflects the natural history of how the cancer evolved. A better exploitation of DNA sequencing would be the use of RNA sequencing to identify which mutated genes are expressed. The advent of NGS has revolutionized transcriptomics and quickly established RNAseq as the preferred methodology for the study of gene expression [101]. RNAseq can also aid in the discovery of novel and unannotated transcripts and identification of SNVs [102]. However, it has been noted that RNAseq is in its infancy and a recent report highlighted that different methodologies do not show high concordance [103].

The standard RNA sequencing method is described as follows: The RNAs in the sample of interest were initially fragmented and reverse transcribed into cDNAs. The cDNAs obtained are then amplified and subjected to NGS. In principle, all NGS technologies can be used for RNAseq such as the Illumina sequencer, which is now the most commonly used platform. The millions of short reads generated can then be mapped onto a reference genome, and the number of reads aligned to each gene gives a digital measure of gene expression levels in the sample under investigation. This helps to identify any differences in sequences such as SNV [101].

6.1.2 Measuring gene expression in cancer

Gene expression is considered to be a key molecular marker for diagnostic and prognostic assessment of cancer [102]. The study of differential gene expression enables the comparison of gene expression profiles from tumour and normal tissues to identify genes that play a major role in the development of the tumour. Here we have used the CLC software to compare the RNAseq from tumour tissues of patients 55, 66 and 73 to the RNAseq of matched normal tissues, to identify differential gene expression profile in these patients. The software used at the time of inception was a 'beta" version, not yet commercially available. As noted above, there is no high quality software that can capture accurately the complete RNAseq landscape (reference). The CLCbio software might have its own limitations, nevertheless, here I describe, for the first time, its use applied to human cancer samples.

6.1.3 Expression of somatic mutation

Somatic mutation calling is traditionally performed on patient matched pairs of tumour and normal genomes/exomes. For a somatic mutation to drive cancer, it must manifest a phenotypic effect. Transcription is the primary conduit by which changes in the genomic code are translated into cellular phenotype [104]. Therefore, the analysis of somatic mutation is better complemented by allelic expression studies, which show whether a mutated allele is expressed. Mutated alleles that are not expressed are unlikely to have any impact on tumour progression [105]. We have analysed the NGS of the whole transcriptome (RNAseq) of three sarcoma samples (55, 66 and 73) to identify genomic-encoded expressed mutations.

6.1.4 Comparing the RNAseq to the DNAseq in CLC

As described in the third chapter, the tumour RNA sequences were compared to the tumour DNA sequences of the same patient. The minimum coverage was set to 5 in the DNA and 1 in the RNA. This comparison results in a list of variants found in the tumour RNA and DNA. We then removed the germline variants found in the normal DNA of the same patient to get a list of expressed somatic mutations only.

6.2 Results and discussion

6.2.1 differentially expressed genes in the tumour and normal tissues of UPS samples

6.2.1.1 Comparing RNAseq of the tumour tissue to the RNAseq of the matched normal tissue in patients 55, 66 and 73

We have used the CLC Biomedical Genomics workbench to identify differentially expressed genes in the tumour and normal tissues of patients 55, 66 and 73. To run the ready-to-use workflow: Toolbox→ ready-to-use workflow→ whole transcriptome sequencing→ human→ identify and annotate differentially expressed genes and pathways.

This analysis results in a list of genes that are overexpressed or suppressed in the tumour of each of the three patients, compared to the matched normal tissue. We then combined the lists of the three patients together into one list to get a mean expression of each gene in the three patients, in normal and tumour tissues. This produces a new list of genes in which overexpressed/suppressed genes in tumour tissues of the three patients are compared to that in their normal tissues. The volcano plot of these genes is shown in figure 6.1.

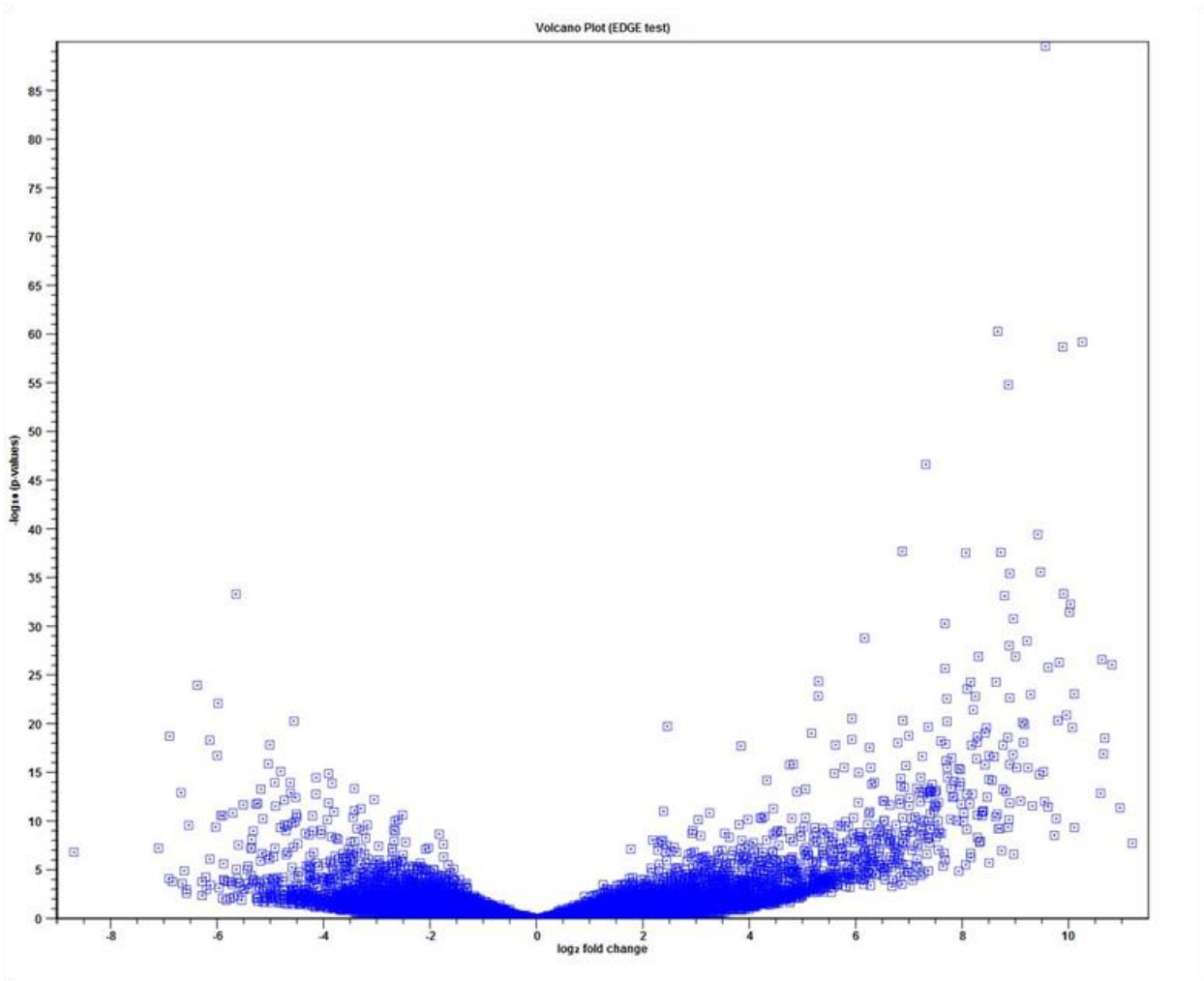


Figure 6.1. Volcano plot of genes that are overexpressed or suppressed in the tumours 55, 66, and 73 compared to the normal tissues of each patient*. Each square represents a gene. The numbers on the X-axis show the log2 fold change, and the number on the Y-axis show the log2 P-values.*

6.2.1.2 Identification the pathways of overexpressed and suppressed genes in the tumour tissues compared to the normal tissues of patients 55, 66 and 73

We have used the *Gene Ontology enrichment analysis and visualization tool* (http://cbl-gorilla.cs.technion.ac.il/) to identify the cellular processes that involve the overexpressed and suppressed genes in the tumour tissues compared to the normal tissues. The processes of overexpressed genes in the tumour with ≥ 5 in fold change are shown in figure 6.2. The main processes include chromatin organization and silencing, nucleosome assembly and organization, protein-DNA complex assembly and regulation of gene expression.

The processes involving the suppressed genes in the tumour (overproduced in the normal tissues), with ≥ 5 in fold change, are shown in figure 6.3. The main processes involve muscle contraction, muscle system processes, muscle cell development and actin-myosin filament sliding. The suppression of these genes may have played an important role in the development of the sarcoma, as it is known that muscles are one of the tissues in which sarcoma can develop.

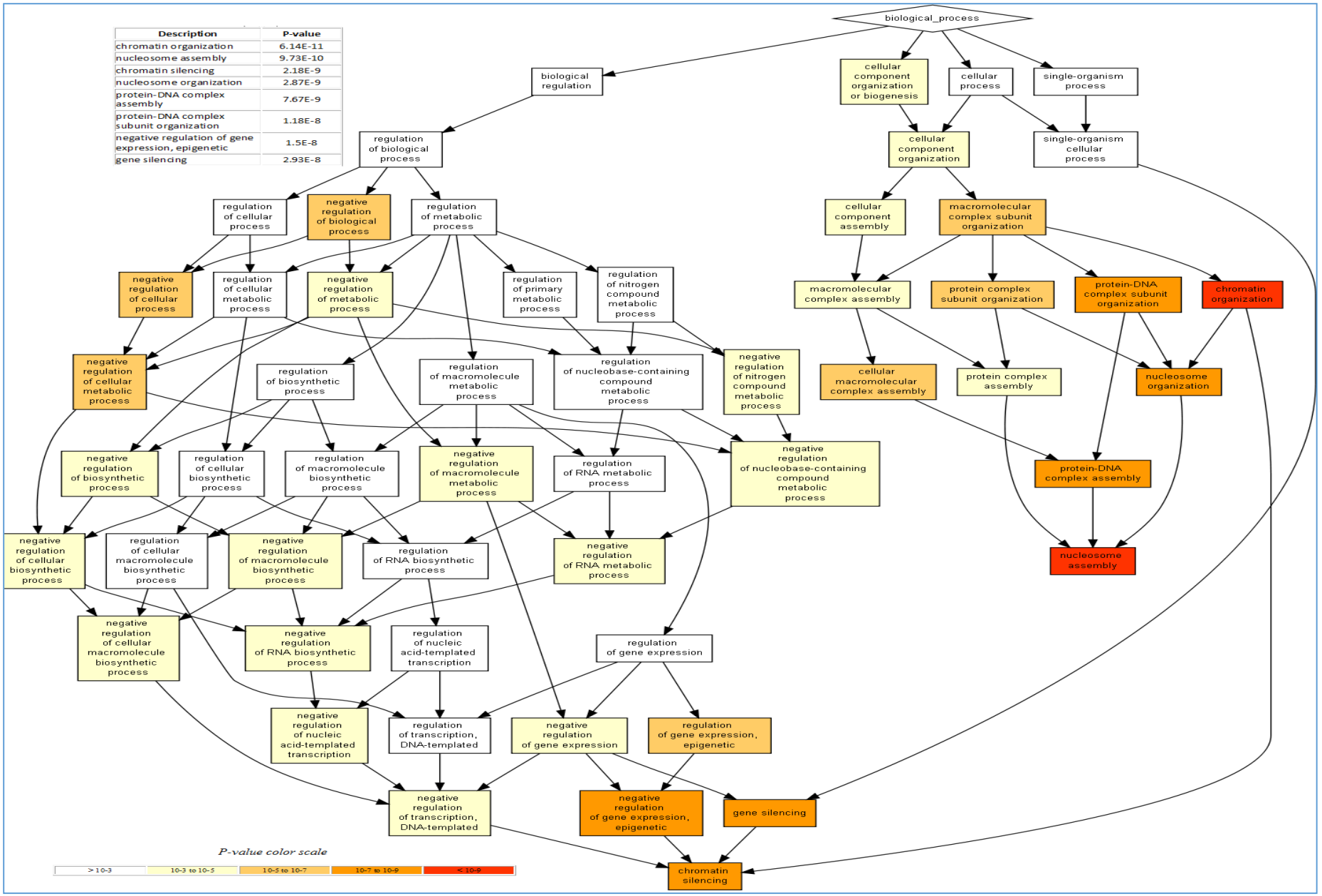
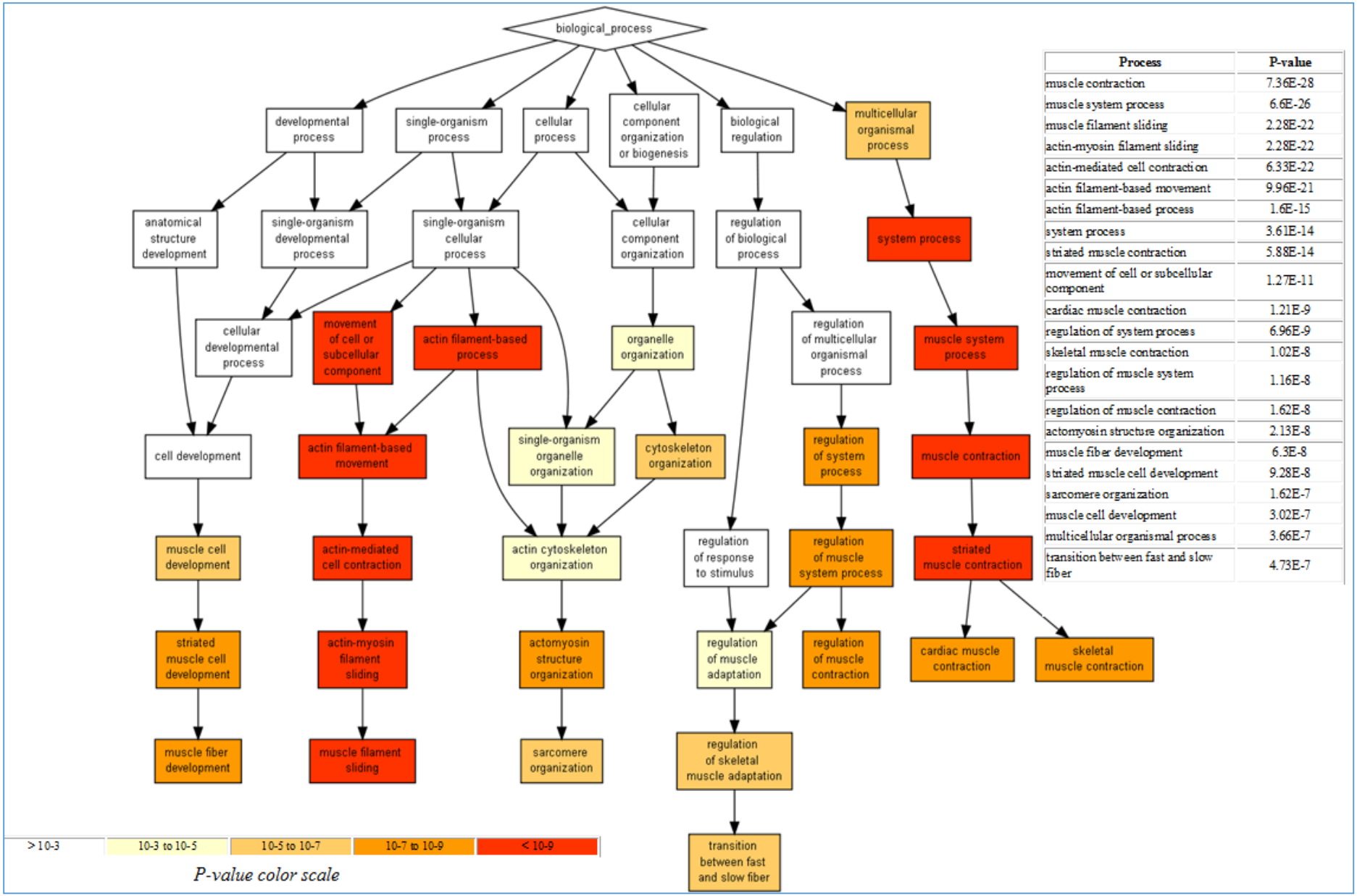

Description | P-value
chromatin organization | 6.14E-11
nucleosome assembly | 9.73E-10
chromatin silencing | 2.18E-9
nucleosome organization | 2.87E-9
protein-DNA complex assembly | 7.67E-9
protein-DNA complex subunit organization | 1.18E-8
negative regulation of gene expression, epigenetic | 1.5E-8
gene silencing | 2.93E-8
biological_process
biological regulation
cellular component organization or biogenesis
cellular process
single-organism process
regulation of biological process
cellular component organization
single-organism cellular process
regulation of cellular process
negative regulation of biological process
regulation of metabolic process
cellular component assembly
macromolecular complex subunit organization
negative regulation of cellular process
regulation of cellular metabolic process
negative regulation of metabolic process
regulation of primary metabolic process
regulation of nitrogen compound metabolic process
macromolecular complex assembly
protein complex subunit organization
protein-DNA complex subunit organization
chromatin organization
negative regulation of cellular metabolic process
regulation of biosynthetic process
regulation of macromolecule metabolic process
regulation of nucleobase-containing compound metabolic process
negative regulation of nitrogen compound metabolic process
cellular macromolecular complex assembly
protein complex assembly
nucleosome organization
negative regulation of biosynthetic process
regulation of cellular biosynthetic process
regulation of macromolecule biosynthetic process
negative regulation of macromolecule metabolic process
regulation of RNA metabolic process
negative regulation of nucleobase-containing compound metabolic process
protein-DNA complex assembly
negative regulation of cellular biosynthetic process
regulation of cellular macromolecule biosynthetic process
negative regulation of macromolecule biosynthetic process
regulation of RNA biosynthetic process
negative regulation of RNA metabolic process
nucleosome assembly
negative regulation of cellular macromolecule biosynthetic process
negative regulation of RNA biosynthetic process
regulation of nucleic acid-templated transcription
regulation of gene expression
negative regulation of nucleic acid-templated transcription
regulation of transcription, DNA-templated
negative regulation of gene expression
regulation of gene expression, epigenetic
negative regulation of transcription, DNA-templated
negative regulation of gene expression, epigenetic
gene silencing
chromatin silencing
P-value color scale
> 10-3
10-3 to 10-5
10-5 to 10-7
10-7 to 10-9
< 10-9

Figure 6.2. Diagram of the main cellular processes of the overexpressed genes detected in UPS by comparing the RNAseq of the tumour tissues of patients 55, 66 and 73 compared to the RNAseq of their normal tissues. *The cellular processes are shown in squares of different colours that represent different p-values, as shown in the p-value colour scale. (http://cbl-gorilla.cs.technion.ac.il/)*



Process	P-value
muscle contraction	7.36E-28
muscle system process	6.6E-26
muscle filament sliding	2.28E-22
actin-myosin filament sliding	2.28E-22
actin-mediated cell contraction	6.33E-22
actin filament-based movement	9.96E-21
actin filament-based process	1.6E-15
system process	3.61E-14
striated muscle contraction	5.88E-14
movement of cell or subcellular component	1.27E-11
cardiac muscle contraction	1.21E-9
regulation of system process	6.96E-9
skeletal muscle contraction	1.02E-8
regulation of muscle system process	1.16E-8
regulation of muscle contraction	1.62E-8
actomyosin structure organization	2.13E-8
muscle fiber development	6.3E-8
striated muscle cell development	9.28E-8
sarcomere organization	1.62E-7
muscle cell development	3.02E-7
multicellular organismal process	3.66E-7
transition between fast and slow fiber	4.73E-7

Figure 6.3. Diagram of the main cellular processes of the suppressed genes in UPS by comparing the RNAseq of the tumour tissues of patients 55, 66 and 73 compared to the RNAseq of their normal tissues*. The cellular processes are shown in squares in different colours that represent different p-values as shown in the p-value colour scale. (http://cbl-gorilla.cs.technion.ac.il/)*

6.2.2 Identification of expressed somatic mutations in RNAseq

6.2.2.1 Number of somatic mutations detected in RNAseq

To look for expressed mutations, we compared the variants between the tumour DNA and the RNA. The 'compare variants in DNA and RNA' ready-to-use workflow identifies variants in DNA and RNA, and studies the relationship between the identified genomic and transcriptomic variants. The important tracks generated were the variants found in both DNA and RNA track, all variants found in DNA or RNA track, and the Genome Browser View.

A low percentage (8.75–17.44%) of somatic mutations detected in the DNA was found to be expressed in the RNA, as shown in table 6.1. Many of the mutated genes might not be expressed at the time that the tumour was obtained, as gene expression is known to be time, cell-type, and stimulus dependent [106]. When the sample is heterozygous for SNV, meaning that there are two different alleles in the same position in the DNA, this may lead to one of two alleles being highly transcribed into mRNA and the other allele being transcribed at either a low level, or not at all. This is known as allele-specific expression [106]. A low coverage or no coverage at the mutation site in the RNAseq is another reason for a detected DNA somatic mutation to be missed in the RNAseq.

A previous study has compared the variants identified by high coverage whole-genome sequencing to those identified by high coverage RNAseq in the same individual, to explore the ability of RNAseq to identify human coding variants [107]. They found that only 40% of exonic variants identified by whole genome sequencing were captured in RNAseq, but this number rose to 81% when concentrating on genes known to be well-expressed in the source tissue.

Sample	Number of mutations detected in TDNA	Number of expressed mutations	Percentage of expressed mutations
55	516	90	17.44%
66	434	38	8.75%
73	399	49	12.28%

Table 6.1. Number of non-synonymous mutations detected in DNA and number of them detected in the RNA of patients 55, 66 and 73

6.2.2.2 Types of expressed somatic mutations detected in RNAseq

The detected somatic mutations in the RNAseq represent all types of mutations detected in the DNA (table 6.2). In patient 55 there were 74 SNVs, one replacement, 6 MNVs, one insertion and 8 deletions. All detected SNVs and MNVs led to amino acid substitutions, except for one SNV in the *HIVEP3* gene that led to a truncation of the protein at amino acid 119. Five deletions and one replacement led to frame-shift mutations (one in *MOV10* gene; one in *FIP1L1* gene; two in *HLA-DQA1* gene; two in *HLA-DRB1* gene), and 3 deletions were in-frame deletions, each with one amino acid deleted, as shown in table 6.2.

In patient number 66, 30 SNVs, 2 MNVs, 2 insertions and 4 deletions were detected. The SNVs and MNVs led to amino acid substitutions, except for one SNV and one MNV, which led to truncation mutations in the *PCDH17* gene at amino acid 1095, and the *CDC27* gene at amino acid 169 respectively (table 6.2). Two of the four deletions were in-frame and deleted one amino acid each, and the other two caused frame-shift mutations in *KIAA0040* gene.

In patient number 73, 30 SNVs, 11 MNVs, 5 insertions and 3 deletions were detected. The 27 SNVs and all MNVs led to amino acid substitutions; 3 SNVs led to truncation mutations; two in the *CDC27* gene and one in the *MAN2C1* gene. Three of the insertions led to frame-shift mutations in *CNOT6L*, *ATP8B3* and *TMEM104* genes. The three deletions were in-frame deletions (table 6.2).

6.2.2.3 Common genes with expressed somatic mutations in the UPS patients

6.2.2.3.1 Common expressed genes in all three patients

NBPF1 (neuroblastoma breakpoint family, member 1) and *CTBP2* genes were mutated in all of the three samples. *NBPF1* has different SNV mutations in each of the three patients, all of them leading to amino acid substitutions; one in patients 55, and 66, and two in patient 73. The normal control coverage is low (only 5) in patient number 55, and only 7 for one mutation in patient number 73 (table 6.2). These mutations with low control coverage need to be validated to make sure they are tumour specific and not false positive variants. When we looked at the detected somatic mutations in the tumour DNAseq, we found that *NBPF1* was detected in all sarcoma samples except in two samples: 59 and 74.

The *CTBP2* gene also has different mutations in each of the three patients: one in 55 and 73, and two in patient 66. They all have low frequency mutations, but are expressed. The normal coverage is high in patient 55 and 66, but is only 12 in patient 73.

NBPF1 is located on 1p36, and this region is frequently deleted in neuroblastoma and other cancer types, including those of neural, epithelial and hematopoietic origin, indicating that this gene may work as a tumour suppressor gene in a broad range of human cancers [108]. A study performed by [108] showed that forced expression of NBPF1 in the human HEK293T cell line resulted in a p53-dependent G1 cell cycle arrest that was accompanied by up-regulation of CDKN1A.

The *CTBP2* gene was detected in 11 other sarcoma samples in the tumour DNAseq. This gene was discussed in chapter 4.

6.2.2.3.2 Common expressed genes in patients 55 and 66

The *SLC7A5* gene was detected as being expressed in patients 55 and 66, with different SNV mutations in the coding region. Again the control coverage is low for both SNVs, with 5 and 8 in patients 55 and 66 respectively. This gene was also detected in four other samples (94, 197, 258 and 469) in their tumour DNA, all with low control coverage.

6.2.2.3.3 Common expressed genes in patients 55 and 73

Three genes were detected in patients 55 and 73: *MTCH2, TPSAB1* and *HLA-DRB1. MTCH2* had four SNVs in patient 55, and one MNV and one SNV in patient 73. All mutations had high coverage in the normal control. *TPSAB1* had one SNV in 55 and one in 73, with high control coverage for both SNVs. *HLA-DRB1* had 3 mutations in patient 55: MNV, replacement and a deletion; and one MNV in patient 73, all with high control coverage.

*MTCH2 (*mitochondrial carrier homologue 2) was also detected in the DNA of patients 59, 84, 94 and 343. The MTCH2 protein is conserved and belongs to the mitochondrial carrier protein family, which catalyses the exchange of solutes across the inner mitochondrial membrane [109]. It was reported that knockout of *MTCH2* in embryonic stem cells and in mouse embryonic fibroblasts inhibited recruitment of tBID (belongs to the Bcl-2 protein family which consists of both pro-apoptotic and anti-apoptotic members) to the mitochondria, and resulted in reduced apoptosis [109].

MTCH2 mutations in patient 55 appear in several reads in tumour DNA, but only in one read in the RNAseq, as shown in figure 6.4A. We tried to validate these mutations in the cDNA of tumour 55 by Sanger sequencing, but failed to detect any of the mutations, as shown in figure 6.4B. This may be due to their low frequency in the RNA.

The *TPSAB1* gene was also detected in the tumour DNA of patients 59, 66, 74, 84, 90, 97A, 197 and 297.

HLA-DRB1 was detected in many other sarcoma samples; 52, 59, 60, 74, 84, 94, 97, 197, 244, 297, 343, 364, 430 and 496. *HLA-DRB1* belongs to the HLA-class II antigens, which are key components of the immune response. They may affect inflammatory responses toward self-antigen and tumour antigen [110].

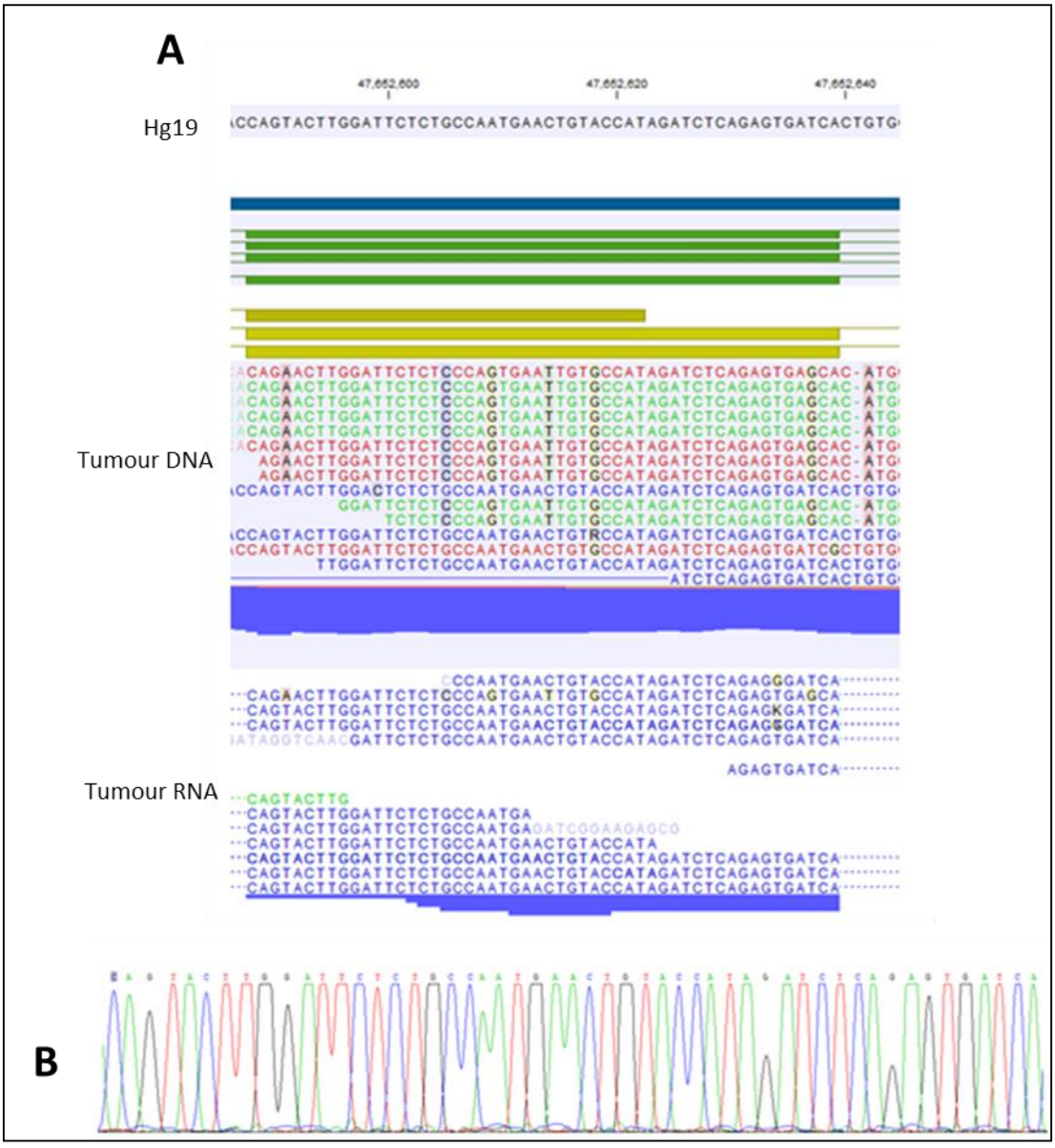


***Figure 6.4. Detection and validation of mutations in MTCH2 gene**. **A,** the CLC detected different somatic mutations in the tumour DNA sequence in patient 55. These mutations were expressed, as they were detected in one read of the tumour RNA seq. Both the DNA and RNA sequences were compared to the normal Hg19 genome sequence. **B,** the cDNA of this region was amplified from the tumour sample of patient 55 and sequenced by Sanger sequence, but failed to detect the variants detected by the CLC.*

6.2.2.3.4 Common genes in patients 66 and 73

There were 5 genes detected with expressed somatic mutations in patients 66 and 73: *CDC27, AHNAK2, CTDSP2, FRG1 and FAT3.*

CDC27 was also detected in the tumour DNA of the patients: 55 (not detected in the RNAseq), 59, 60, 74, 84, 94, 97, 297, 364, 430 and 496. CDC27 is a cell cycle regulator, which participates in control of the mitotic checkpoint and surveys the mitotic spindle to maintain chromosomal integrity [111]. Therefore, mutations in this gene may disturb its function and influence the mitotic progression of cells. One study has shown that the expression of CDC27 was down-regulated in breast cancer cell lines and postulated that it may act as a tumour suppressor [111].

AHNAK2 was also detected in patients 52, 55, 60, 84, 90, 343 and 378. *CTDSP2* was detected in 59, 90, 97B, 197, 343 and 430. FRG1 was detected in patients 52, 97B, 258, 297, 343, 364, 378 and 496.

The *FAT3* gene is only detected in patients 66 and 73 (fig 6.5). However, other FAT genes are detected with somatic mutations in the DNA in other patients: *FAT4* in 52 and 378, *FAT1* in 90, *FAT2* in 297. All of them have been reported to be mutated in several cancer types and believed to work as tumour suppressors [112].

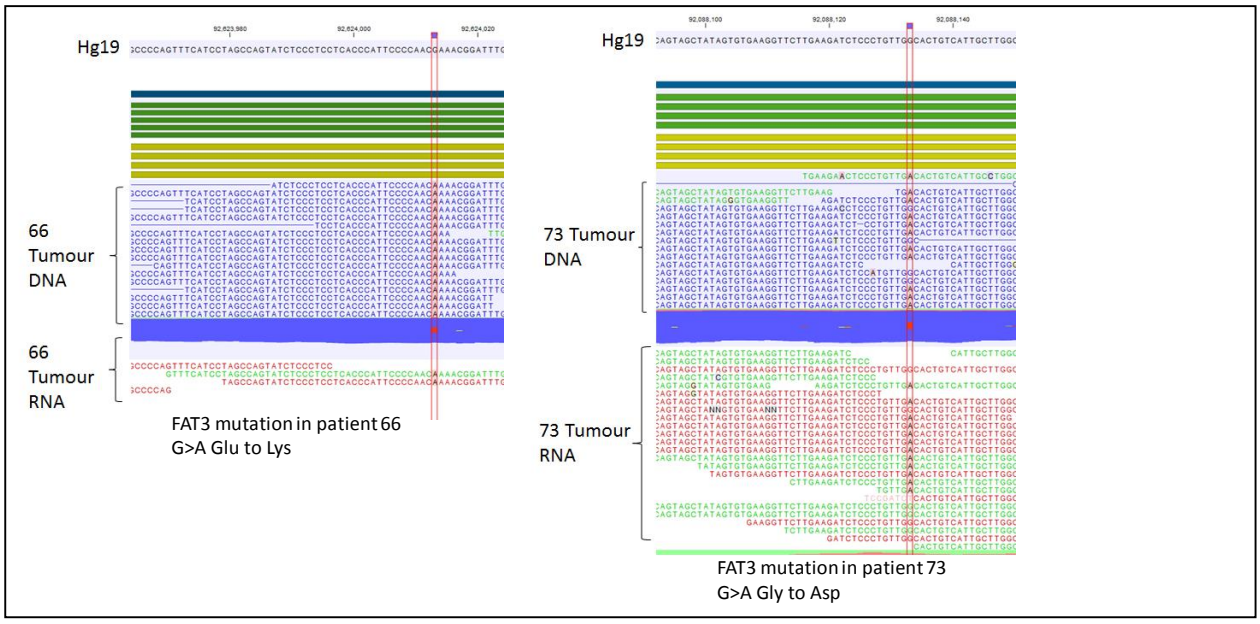


Figure 6.5. FAT3 RNA expressed somatic mutations in patients 66 and 73. *The CLC browser of FAT3 mutations in the tumour DNA and tumour RNA of patients 66 and 73. The sequences are compared to the Hg19 sequence*.

6.2.3 Individualised mass spectrometric (MS) proteomics using the patient specific mutant references databases to identify expressed tumour specific antigens

6.2.3.1 Proteogenomics

As stated above, the genome DNA sequence can be considered to be static. The expressed cancer genome (e.g. RNA and protein expression) more accurately describes the cancer samples at the time of presentation in the clinic. It would also be of interest to compare the mutated cancer genome with the expressed RNA landscape and the proteome to define how this information, in total, can be used to describe the cancer tissue in molecular terms. Improvements in mass spectrometry (MS)-based peptide sequencing provide new opportunities for determining whether or not a somatic mutation is translated. Detection of somatic mutations by NGS of cancer genomes and transcriptomes has demonstrated their enormous complexity, and it is often unclear which somatic mutations drive tumour biology and which are non-functional passenger mutations. Mutation detection at the peptide level clearly increases the confidence that any given variant is potential biological driver. Thus, integrated proteogenomic methods that combine the DNA/RNA sequences and proteomics are of particular importance for identifying novel peptides resulting from somatic mutations [113]. Determination of which DNA or RNA sequence level variants are expressed as proteins provides a basis for prioritizing mutations for further study of their contributions to cancer phenotypes [114].

In general, proteins extracted from each sample (tissue) are digested using enzymes with a high specificity for known cleavage sites, such as trypsin. Then, tandem mass spectra are generated for individual peptides using mass spectrometry. These peptide spectra can then be identified, by searching for matches in generalized human proteome databases [115].

6.2.3.2 Proteomics of patient 55

Protein lysate from patient 55 (tumour vs normal tissue) was processed using MS to identify total proteins present in the tumour (performed as a collaboration between the Hupp laboratory and Dr Borek Vojtesek's laboratory, Brno, Czech Republic). The proteins overproduced in the tumour vs normal adjacent tissue are shown in Figure 6.6 (scatter plot).

A recent study has carried out a proteome screen using tandem mass tag isobaric labelling on three high-grade UPS, one of which is the tumour 55, and has identified the commonly dysregulated proteins within the three sarcomas. It has further validated the most penetrant

receptors, using immunohistochemistry (ICH), arising from two different population cohorts representing over 300 patients [116]. The study has shown that there were 749 proteins expressed in the three UPS biopsies, but it focused on dominant transmembrane receptors that might be suitable as targets for receptor inhibitors in the future, such as CD44 and CLIC1. The study showed that CD63, CD44, CLIC1 and CLIC4 were overexpressed in the three sarcomas, compared to the normal tissue 22 (Figure 6.7), and confirmed their elevated expression by qRT-PCR and ICH. CD44 is a cell-surface glycoprotein that is known to play roles in extracellular cell–cell contacts, cell adhesion and migration, and it has been shown that its absence prevents metastasis in osteosarcoma in mice [116]. CLIC1 was shown to work as a pro-metastatic factor in a set of cancer models [116]. Interestingly, CD44, CLIC1 and CLIC4 were detected as overexpressed proteins in tumour 55, compared to normal tissue, with fold changes of 89.2, 73.02, and 56.04 respectively. CD44 and CLIC4 were also detected in the overexpressed genes in tumour 55 generated by RNAseq. This validates patient 55 tumour as “sarcoma pleomorphic” at the proteome level.

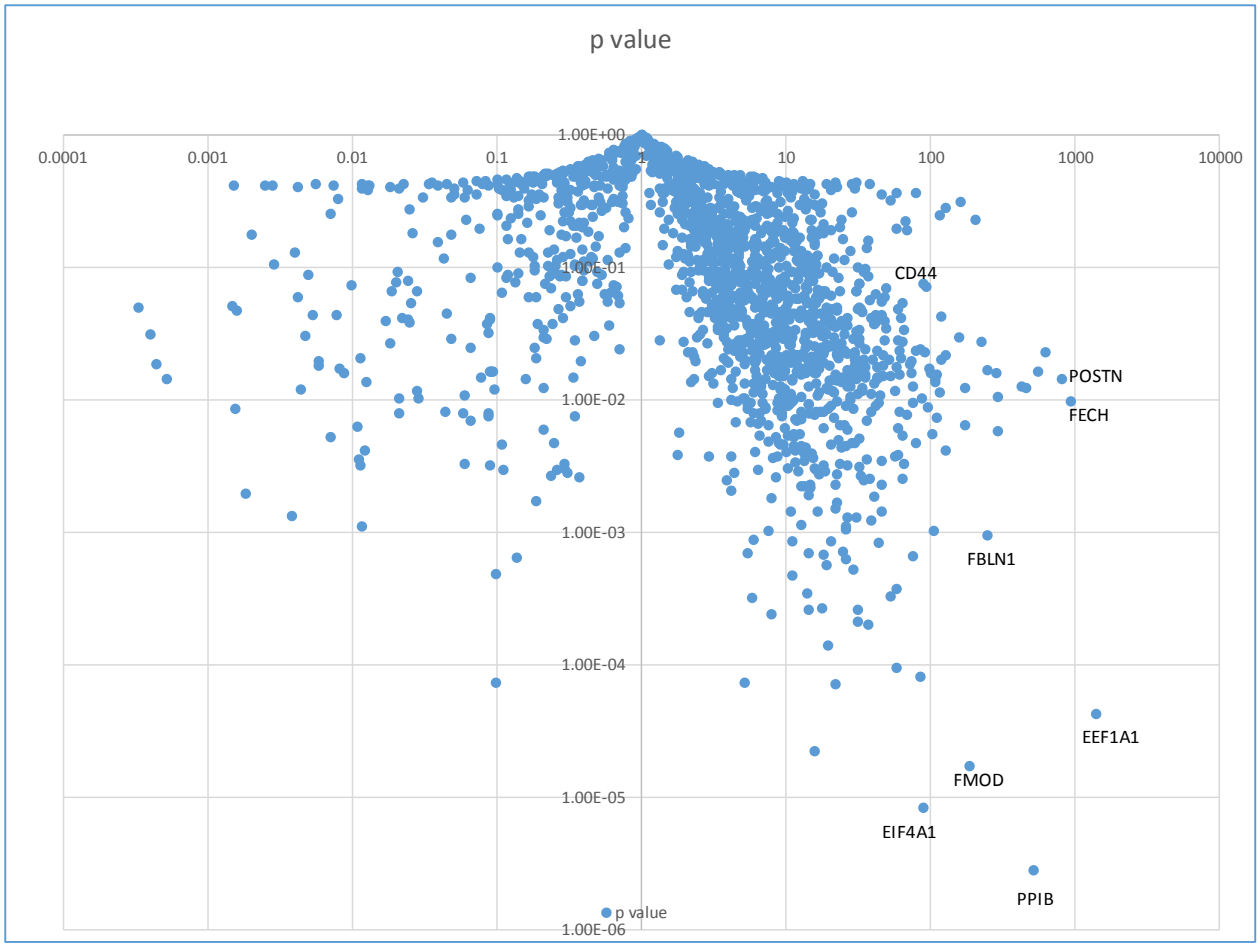


Figure 6.6. Scatter plot of the differentially expressed proteins in the tumour 55 compared to the normal tissue generated by MS*. Each blue dot represents a protein with specific p-value (Y-axis) and fold change (X-axis). Some of the highly produced proteins are named in the figure*.

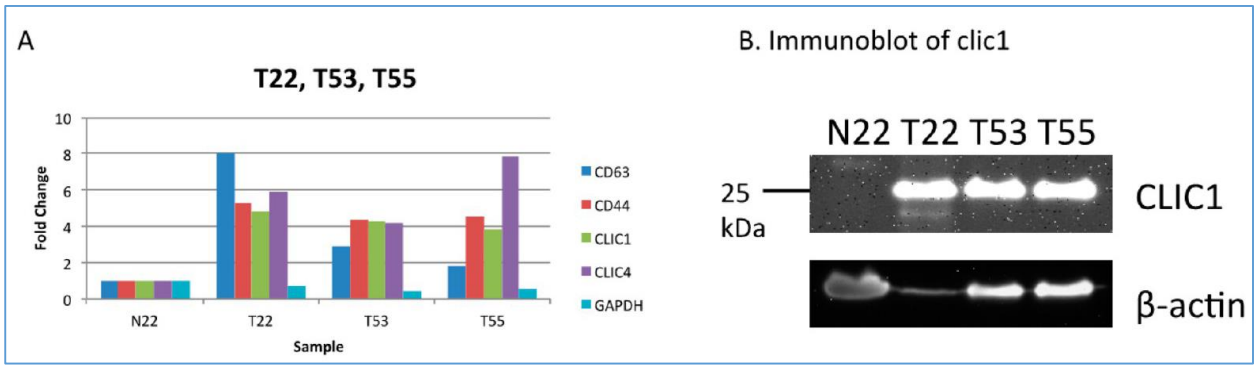


Figure 6.7. Differential expression of CD44, CD63, CLIC1, and CLIC4 in three sarcoma patients. A, *Summary of quantitative protein expression of the four proteins CD44, CD63, CLIC1, and CLIC4 expression in T22, T53, and T55, compared with each other and with N22, are shown with GAPDH included for comparison as a control.* ***B,*** *of these proteins, CLIC1 is the most novel pro-oncogenic target, and as such we determined whether standard immunoblotting could be used to confirm the TMT data. Lysates from all three cancers (T22, T53, and T55) and the normal control (N22) were immunoblotted, using antibodies to CLIC1 and included actin as a loading control [116].*

6.2.3.3 Overexpressed genes/proteins in the RNAseq list and proteins list

We compared the highly expressed proteins (≥ 1.5-fold change) in tumour 55 with the overexpressed genes detected by RNAseq in the three sarcoma patients to see if there are any common genes in the two lists. We found 163 proteins, which are overproduced in tumour 55, that are also detected in the overexpressed genes (≥ 5-fold change) by RNAseq. These genes are listed in table 6.3.

Number	Gene Name	Number	Gene Name	Number	Gene Name
1	*ACAN*	60	*FARP1*	119	*RAC1*
2	*FBN2*	61	*PLOD1*	120	*IGF2BP2*
3	*HIST1H1B*	62	*AGRN*	121	*RHOC*
4	*POSTN*	63	*CTSB*	122	*RPS2*
5	*CRABP1*	64	*PDXK*	123	*IL4I1*
6	*TNC*	65	*CALD1*	124	*ALDH1A3*
7	*IGFBP2*	66	*EIF4A1*	125	*GSN*
8	*MDK*	67	*GLB1*	126	*FBLN1*
9	*AEBP1*	68	*PLOD2*	127	*XPNPEP1*
10	*PTK7*	69	*CLIC4*	128	*RPS3*
11	*MFAP2*	70	*ANPEP*	129	*STARD5*
12	*COL16A1*	71	*YARS*	130	*RECQL*
13	*FN1*	72	*CTSA*	131	*FNDC3B*
14	*MMP14*	73	*ANXA2*	132	*LGALS1*
15	*CD276*	74	*GPX8*	133	*MRC2*
16	*MEST*	75	*ITGB1*	134	*BAX*
17	*ANO1*	76	*ISG15*	135	*TXNDC5*
18	*COL6A3*	77	*SCRN1*	136	*COPA*
19	*HIST1H1D*	78	*MCM4*	137	*NUP50*
20	*TFRC*	79	*OLFML3*	138	*A2M*
21	*IGF2*	80	*PALLD*	139	*IFI30*
22	*TNMD*	81	*SEC24D*	140	*NID2*
23	*TGFBI*	82	*MXRA5*	141	*VGLL4*
24	*LEPRE1*	83	*PRDX4*	142	*COL4A2*
25	*HIST1H4A*	84	*GMFG*	143	*TAGLN2*
26	*COL18A1*	85	*THY1*	144	*AKAP2*
27	*SPRY4*	86	*AKAP13*	145	*PDIA5*
28	*HAPLN3*	87	*ARHGDIA*	146	*PPIA*
29	*COL12A1*	88	*SCARB2*	147	*SERPINB9*
30	*SERPINE2*	89	*SEC23A*	148	*SERPINA1*
31	*GPNMB*	90	*VKORC1*	149	*UPF1*
32	*CRABP2*	91	*LAMB1*	150	*ARPC1B*
33	*ISLR*	92	*VCAN*	151	*LGALS3BP*
34	*FSCN1*	93	*BASP1*	152	*LTBP4*
35	*LOXL2*	94	*P4HB*	153	*SEC13*
36	*HIST1H2BN*	95	*FKBP14*	154	*IFI16*
37	*ZNF185*	96	*COL6A2*	155	*ENO1*
38	*CDKN2A*	97	*GGH*	156	*PPP2R4*
39	*SERPINH1*	98	*MX1*	157	*PDIA3*
40	*TIMP1*	99	*ARMCX2*	158	*NCL*
41	*FKBP10*	100	*ARF4*	159	*IQGAP1*

42	*CTNNB1*	101	*CSRP2*	160	*NDRG1*
43	*LTBP1*	102	*MAN1B1*	161	*CD44*
44	*GUSB*	103	*CADM1*	162	*GANAB*
45	*HIST1H1A*	104	*S100A4*	163	*HDAC1*
46	*COL5A1*	105	*PAFAH1B3*		
47	*GARS*	106	*RAB31*		
48	*PXDN*	107	*CASK*		
49	*PDGFRB*	108	*ANXA5*		
50	*HM13*	109	*FMOD*		
51	*CALU*	110	*SPARC*		
52	*FLNA*	111	*PRKCSH*		
53	*LEPREL2*	112	*GALE*		
54	*MFGE8*	113	*PHGDH*		
55	*PCOLCE*	114	*STMN1*		
56	*COL6A1*	115	*PPIB*		
57	*SEC31A*	116	*RCN3*		
58	*CNN3*	117	*POFUT2*		
59	*EMILIN1*	118	*ISYNA1*		

Table 6.3. Shared overexpressed genes/proteins in the differentially expressed gene profiles generated by RNAseq, and the overproduced proteins detected by MS in patient 55.

Some of these highly expressed proteins have been reported to be elevated in other cancers such as POSTN and FMOD. The periostin (POSTN) gene encodes the ligand for integrin, one of the key focal adhesion proteins contributing to the formation of a structural link between the extracellular matrix and integrins. High expression levels of the POSTN gene are correlated with numerous human malignancies [117]. It was reported that POSTN over-expression in colorectal cancer was positively correlated with tumour size, differentiation, lymph node metastasis, clinical stage and five-year survival rate [117]. FMOD (fibromodulin) was found overexpressed in virtually all patients with chronic lymphocytic leukaemia tested, and not detected in circulating B cells and other peripheral blood cells, which suggests its important role in tumourigenesis [118].

6.2.3.4 Identification of mutant proteins using the genomic data

In addition to total proteomic analysis to identify pathways elevated in tumours, emerging aims include identifying mutated proteins, which requires bespoke reference databases from patients' tumours. These mutated proteins can be drivers, and may be used as potential vaccines via MHC Class I neoantigen presentation. To determine whether our genomics (DNAseq and RNAseq) can be used to detect mutated peptides, we formed a mutant reference database of over 300 potentially mutated proteins derived from the DNA and RNAseq analysis.

Specifically, to identify mutated peptides using MS, we cross-referenced >4,000 proteins detected in the tumour to the variants (indels and SNVs; patient 55) to create a list of possible mutated proteins. We then created an *in silico* tryptic library of mutated peptides based on the DNA variant list (data not shown). This library was used to search mass spectra files for tryptic peptides with a mutation. Using this approach, we could detect >25 mutant peptides in the tumour. As a specific validation to highlight, the *PDCD6* gene was found mutated in the tumour, expressed RNA was detected with a mutation, and the mass spectra was able to identify the mutated tryptic peptide (V98I) (Fig 6.8). SWATH-MS also showed that PDCD6 was elevated. Together, these data highlight the workflow that can be used to identify expressed mutant peptides, based on DNAseq, RNAseq, and MS.

Figure 6.8. Mutant peptide identification. ***A,*** *the browser shows a PDCD6 mutation from cancer DNA using CLCbio software (top; "A" mutation). Orthogonal validation is achieved using RNAseq (bottom; "A" mutant mRNA).* ***B****, uploading over 100 mutant tryptic sequences into Bruker software allows identification of the mutant peptide with fragmentation shown.* ***C****, the mutant peptides thus identified can form a database from which a viral based vaccine can be generated.*

6.2.4 RNA editing

6.2.4.1 Number of variants specific to RNAseq

As stated above, software used to annotate RNA expression is in its infancy and there are no "gold standards" yet available for widespread use. Our collaborators at CLCbio have developed their software to be used as a tool for RNA editing events. When we compared the RNAseq to the DNAseq of each patient, the CLCbio software detected many variants in the RNAseq that are not detected in the matched DNA of the same patient. There were 4118, 3112 and 3304 specific RNA variants detected in patients 55, 66 and 73 respectively that are not encoded by the DNA. This is suggestive of RNA editing.

6.2.4.2 Main types of RNA editing

RNA editing is the changing of nucleotide sequence of RNA transcripts relative to that of the encoding DNA [119]. There are two major types of RNA editing in mammals: adenosine to inosine (A to I), and cytidine to uracil (C to U). The A to I editing is catalysed by the ADAR (adenosine deaminase acting on RNA) class of enzymes that bind double-stranded RNA [120]. In fact, the base I is read as guanosine and it does not pair with U but instead with C [121]. The C to U editing is catalysed by a diverse family of 10 different cytidine deaminases, called APOBECs, that can target RNA and DNA. The effects of RNA editing by ADAR and APOBECs are summarised in figure 6.9.

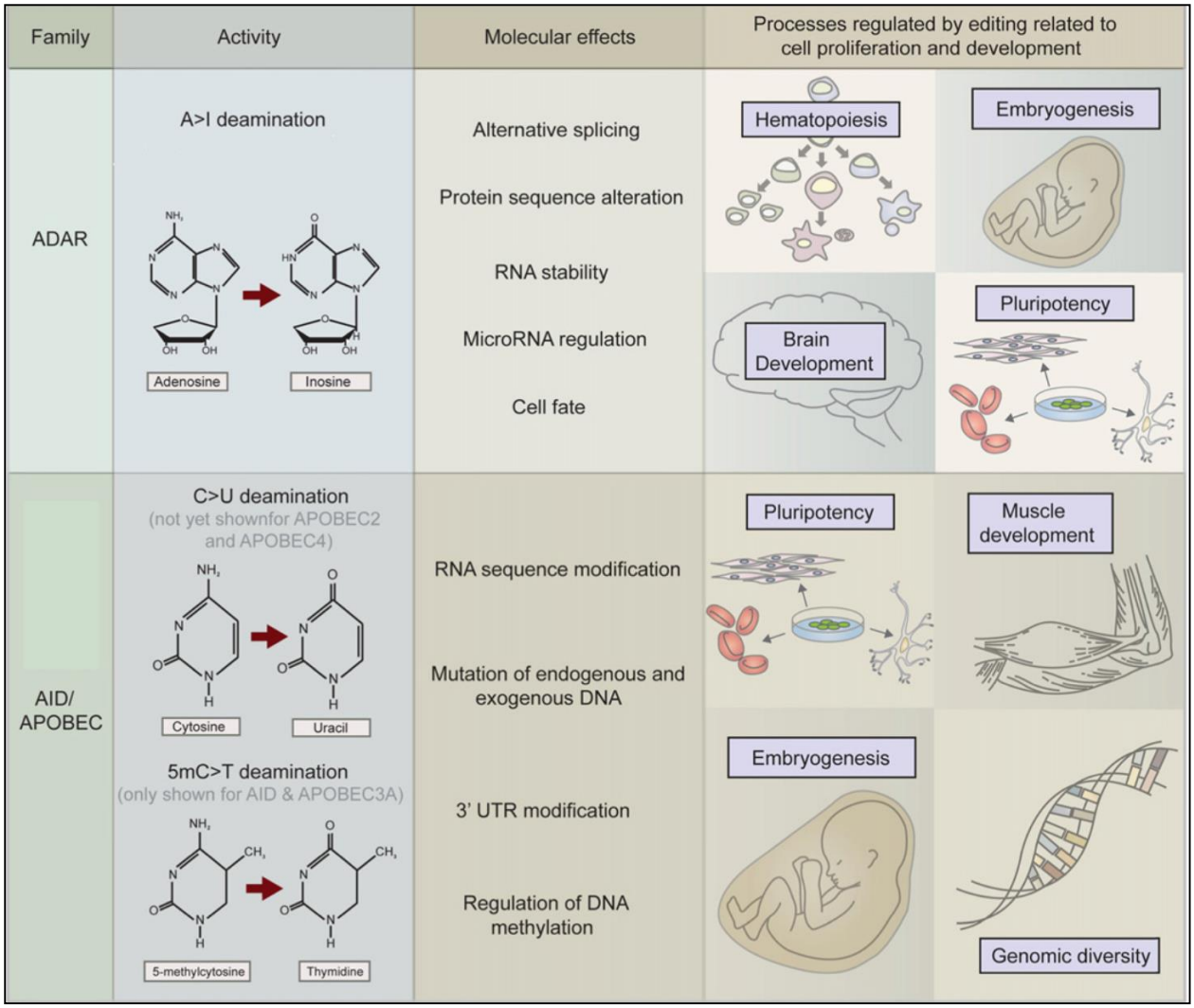


***Figure 6.9. Effects of RNA editing by ADAR and APOBEC enzymes**. Base deamination is carried out by ADAR (A to I) and AID/APOBEC (C to U; C to T) protein families that result in extensive molecular effects such as ADAR: alternative splicing, protein sequence alteration, RNA stability, miRNA regulation and cell fate; APOBEC: RNA sequence modification, DNA mutation, 3′ UTR modification and regulation of DNA methylation. Furthermore, editing changes have widespread effects but many affect processes that are related to cell proliferation and development, e g., ADAR: haematopoiesis, embryogenesis, brain development and pluripotency, APOBEC: muscle development, pluripotency, embryogenesis and genomic diversity [122].*

6.2.4.3 RNA editing effects and its role in cancer

RNA editing is believed to be regulated, possibly in response to cellular or extracellular stimuli, and can affect several steps during gene expression and regulation, such as splicing, RNA stability, localisation and translation [120].

A significant emerging consequence of abnormal editing is a strong connection to cancer biology. Associations with both solid tumours and blood cancers have been observed for multiple ADAR and APOBEC family members [122]. RNA editing can modulate the expression of oncogenes or tumour suppressor genes, for example by causing non-synonymous changes to coding regions, which may act as drivers for tumour growth and serve as prognostic or

predictive markers for patient stratification [123]. The data from differential gene expression profiles generated from RNAseq of tumour and normal tissues show that there is elevated expression of different ADAR and APOBEC transcripts in the tumours of patients 55, 66 and 73 (fig 6.10), which may explain the high rate of RNA editing in these tumours.

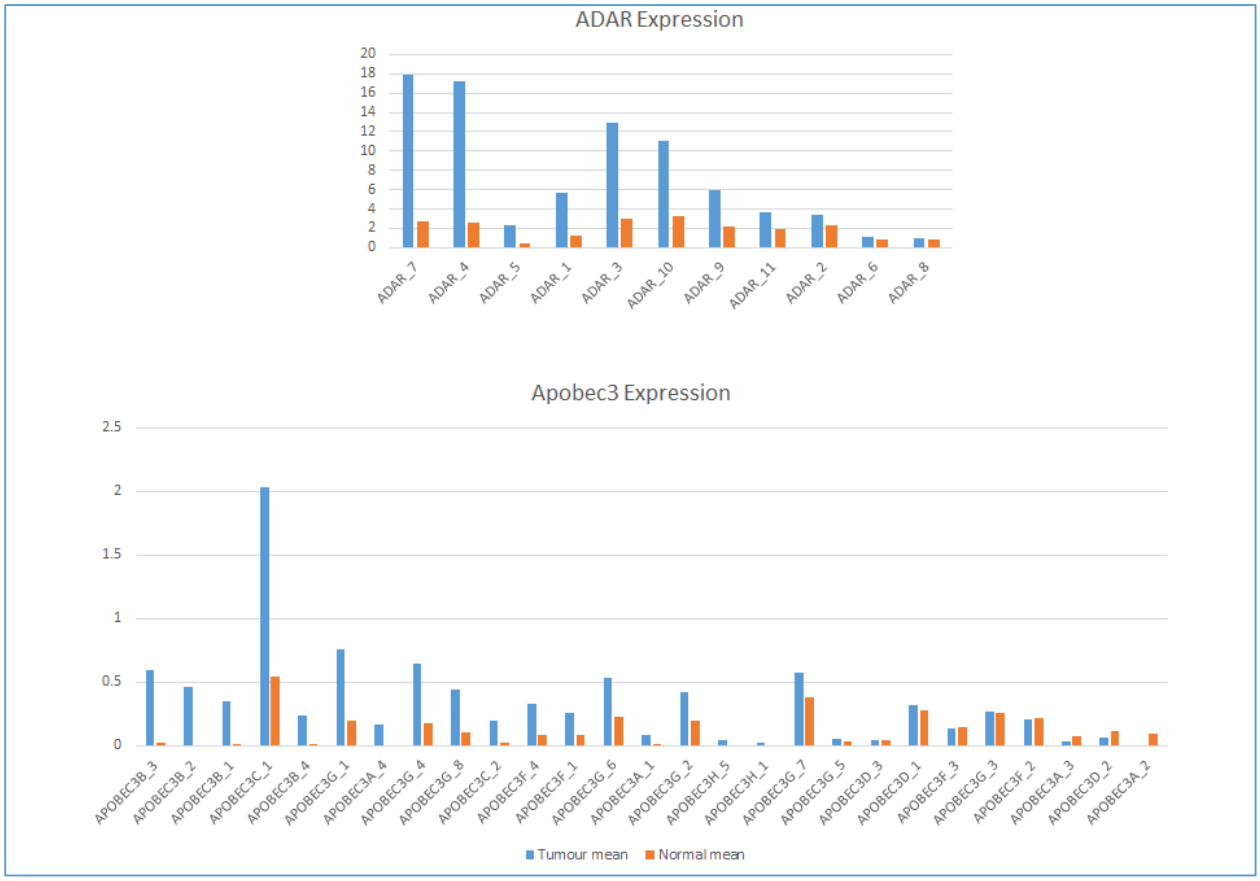


Fig 6.10. Overexpression of ADAR and APOBEC transcripts in the tumours of patients 55, 66 and 73*. The mean expression levels of ADAR and APOBEC transcripts in the tumour RNAseq (blue) and normal RNAseq (brown) in patients 55, 66 and 73 are shown.*

6.2.4.4 RNA editing in the *MAP4K5* gene

One example of the detected RNA editing is A to T in the second transcript of the gene *MAP4K5,* as shown in figure 6.11. The CLCbio has detected this edit in the RNAseq of patient 73 with high frequency (69%). The edit was validated by Sanger sequence as shown in figure 6.11C. The edit changes the amino acid valine at position 569 to glutamic acid. It is at the intron–exon junction and seems to be important for splicing. As shown in figure 6.11(A, C), the first transcript of the gene does not have the codon CTA that has the edit, and this transcript has the exon that has R1 primer in the figure 6.3A, whereas the second transcript, which has the CTA codon with the edit, skipped the exon with the R1 primer. The edit was

also confirmed to occur in the normal RNA with a lower frequency in all sarcomas tested (data not shown).

In another experiment, the whole exome and transcriptome of the A375 cell line was sequenced and analysed in the same manner as the sarcoma samples, to validate the RNA-editing software. We were able to recapitulate some of the RNA edits observed in clinical tumour tissue, such as the edit found in the *MAP4K5* gene (performed by Giuse Grasso in the Hupp laboratory). This edit merits further studies to see if it has a role in tumourigenesis.

In a recent study [124] the expression of MAP4K5 in pancreatic ductal adenocarcinoma (PDAC) was examined. It showed that MAP4K5 expression was decreased or lost in 77.1% of PDAC, but it was expressed at a high level in the normal pancreatic ductal cells in 100% of the matched non-neoplastic pancreas samples. It also showed that the loss of MAP4K5 expression was associated with reduced overall survival of pancreatic cancer patients. Thus, it is believed that MAP4K5 works as a tumour suppressor and its loss contributes to the development of the tumour.

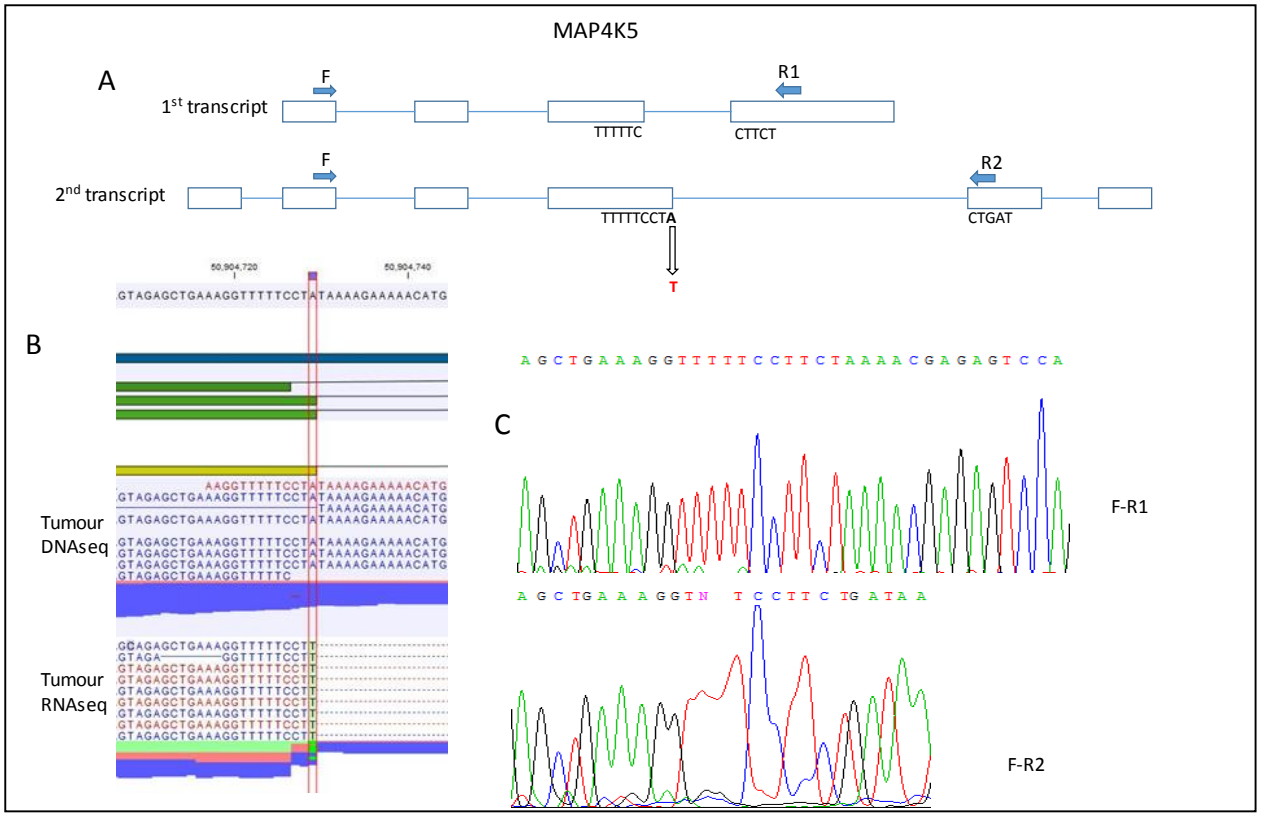
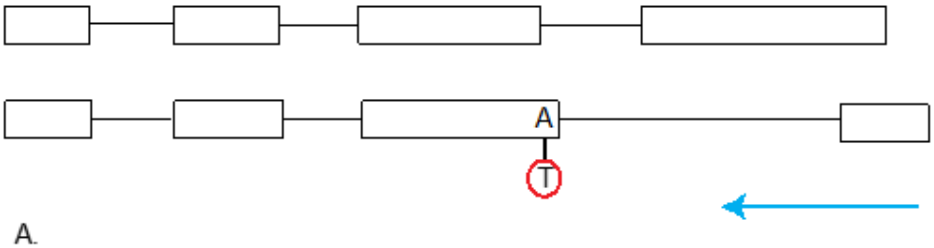


Figure 6.11. Detection of RNA editing (A>T) in MAP4K5 gene. ***(A)*** *Diagram shows two transcripts of the MAP4K5 gene. The edit location is shown in the 2nd transcript as A to T. the location of the primers used to amplify the regions are shown by F1, F2 and R.* ***(B)*** *The CLC browser shows the edit in the RNAseq of patient 73 and the normal sequence in the tumour DNA match and Hg19.* ***(C)*** *The Sanger sequence validation of the 1st and 2nd transcript amplified by the primers (F-R1) above and (F-R2) below.*

6.2.4.5 MAP4K5 RNA editing in different species

When we looked at the human MAP4k5 protein in Ensembl; we found that there are three different transcripts in regards to the codon that has the RNA edit (fig 6.12 A). The first one has no edit in the codon GTA which codes for the Valine amino acid. The second transcript has the edit in the codon and becomes GAA which codes for the amino acid Glutamic acid. The third transcript has this codon deleted and the so the amino acid. Pan species has the same three transcripts to humans (fig 6.12 B). Murine species has the codon GCA, instead of GTA, in one transcript which codes for Alanine. The other two transcripts are similar to the second and third transcripts to humans with the edited codon GAA and with the deleted codon respectively (fig 6.12 C). the Danio species have two transcripts only; one with the edited codon GAA, and one with the deleted codon (fig 6.12 D). so this edit is conserved in many species and seems to be important for alternative splicing to produce different forms of MAP4K5 protein.



A.
Human splice form α
M S L S V G K T F
-ATGTCATTATCAG--GTATGTATGT------TTTTTCTTTTA--TAGGAAAAACCTTT-

Human splice isoform β
M S L S E G K T F
-ATGTCATTATCAG--GTATGTATGT------TTTTTCTTTTA--AAGGAAAAACCTTT-

Human normal isoform χ
M S L S G K T F
-ATGTCATTATCAG--GTATGTATGT------TTTTTCTTTTATAG--GAAAAACCTTT-

B.
Pan MAP4K5-201 ENSPTRT00000011589.5 splice form α
M S L S V G K T F-
-ATG-TCA-TTA-TCA-GTA-GGA-AAA-ACC-TTT-

Pan MAP4K5 XP_001155473.1 splice form β
M S L S E G K T F-
-ATG-TCA-TTA-TCA-GAA-GGA-AAA-ACC-TTT-

Pan MAP4K5 XM_001155473.5 splice form χ
M S L S G K T F-
-ATG-TCA-TTA-TCA-GGA-AAA-ACC-TTT-

C.
Murine MAP4K5-001 ENSMUST00000110570.7 splice form α
M S L S A G K T F-
ATG-TCA-TTA-TCA-GCA-GGA-AAA-ACC-TTT

Murine MAP4K5-201 ENSMUST00000049239.7 splice form β
M S L S E G K T F-
ATG-TCA-TTA-TCA-GAA-GGA-AAA-ACC-TTT

Murine MAP4K5 XP_006516117.1 splice form χ
M S L S G K T F-
-ATG-TCA-TTA-TCA-GGA-AAA-ACC-TTT-

D.
Danio MAP4K5-001 ENSDART00000029824.7 splice form α
M S L S G K T F-
-ATG-TCA-TTA-TCA-GGG-AAG-ACG-TTT

Danio MAP4K5 XP_017214192.1 splice form β
M S L S E G K T F-
ATG-TCA-TTA-TCA-GAA-GGG-AAG-ACG-TTT

***Figure 6.12. RNA editing and different splice forms of MAP4K5 in four different species. A and B**, there are three different transcripts: no edit in the GTA codon which codes for Valine, edited to become GAA which codes for Glutamic acid, and the codon is deleted in humans and pans respectively. **C,** Murine has GCA codon in one form which codes for Alanine, the other two forms are similar to the humans and pans 2nd and 3rd forms. **D,** Danio has only two forms: one with the edited codon GAA and one with the deleted codon.*

6.2.4.6 Expression of MAP4K5 in UPS

We have performed PCR to amplify part of the 2nd transcript cDNA, with the primer locations shown in figure 6.3A, from different sarcoma samples and their matched normal samples. The PCR results indicate that there is more expression of the MAP4K5 transcript in the tumours T55, T73, T52 and T97 than the normal tissues. This appeared as larger bands in the agarose gel as shown in figure 6.13A. This is also supported by RNAseq data that shows overexpression of MAP4K5 gene transcripts in the tumour of patients 55, 66 and 73 compared to the tumour tissues of these patients (fig 6.13B)

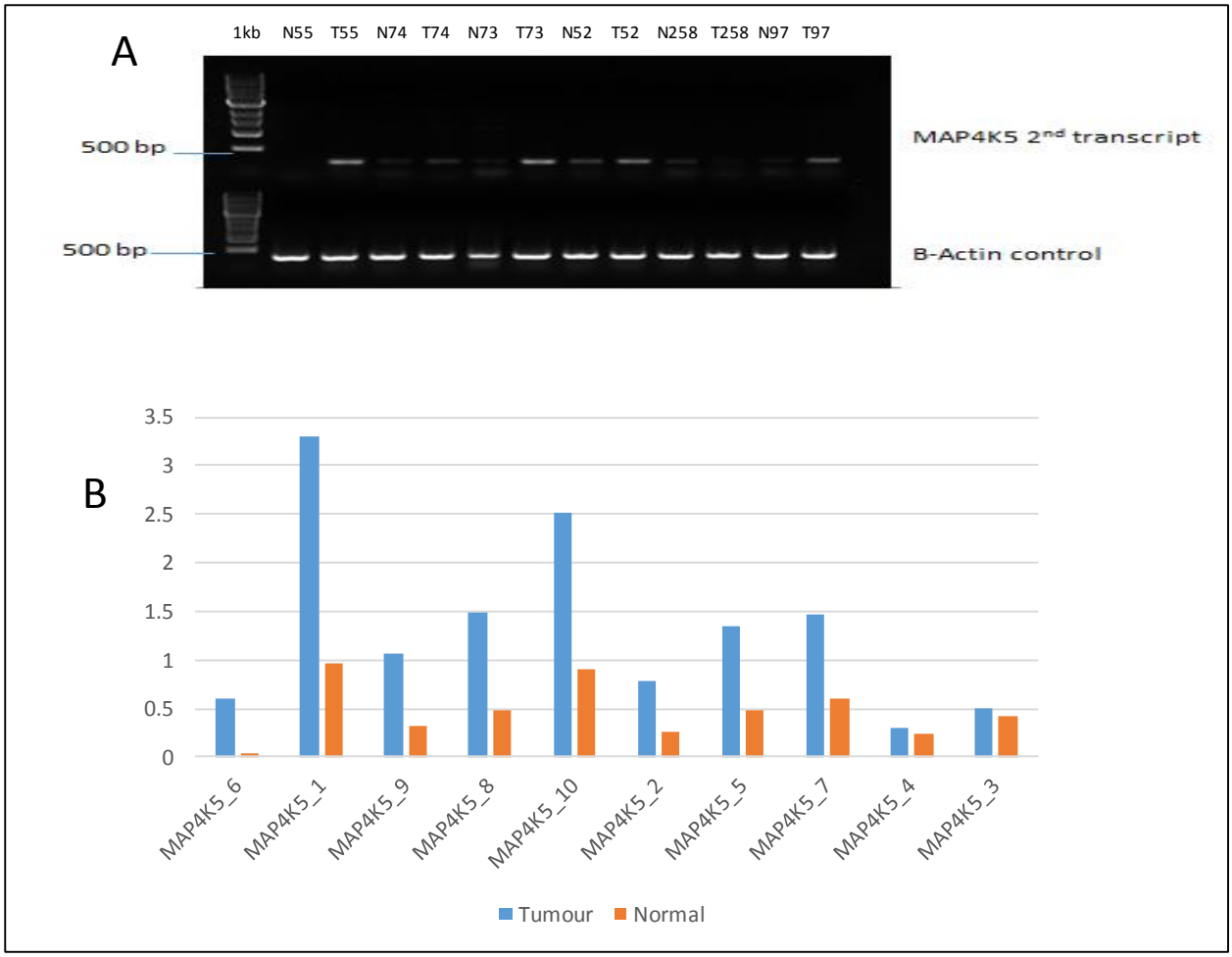


Figure 6.13. Overexpression of MAP4K5 gene in the tumour tissues. A, *Gel electrophoresis image of the PCR results of the 2nd transcript of MAP4K5 gene shown in figure 5.3A. The F2-R primers were used to amplify the 260 bases of MAP4K5 gene from the cDNA which is made from the RNA from the tumour and normal tissues of the sarcoma patients 55, 74, 73, 52, 258 and 97. Beta actin primers were used as a control.* ***B,*** *Mean expression of MAP4K5 gene transcripts in the tumour (blue) and normal (brown) tissue, in the RNAseq of patients 55, 66 and 73.*

6.2.4.7 RNA editing specific to tumour RNA

We found that some of the detected RNA editing variants were also present in the normal RNA, such as the edit in *MAP4K5*. Therefore, we followed the steps shown in figure 6.14 to restrict the detected variants to the exome regions of the tumour RNAseq. This results in many SNVs specific to tumour RNA and each variant has coverage of at least 5 in the DNAseq. There were 3397 variants detected in patient 55; 2194 of them were non-synonymous, 1463 variants in patient 66; 1033 of them were non-synonymous, and 1367 variants in patient 73; 971 of them were non-synonymous.

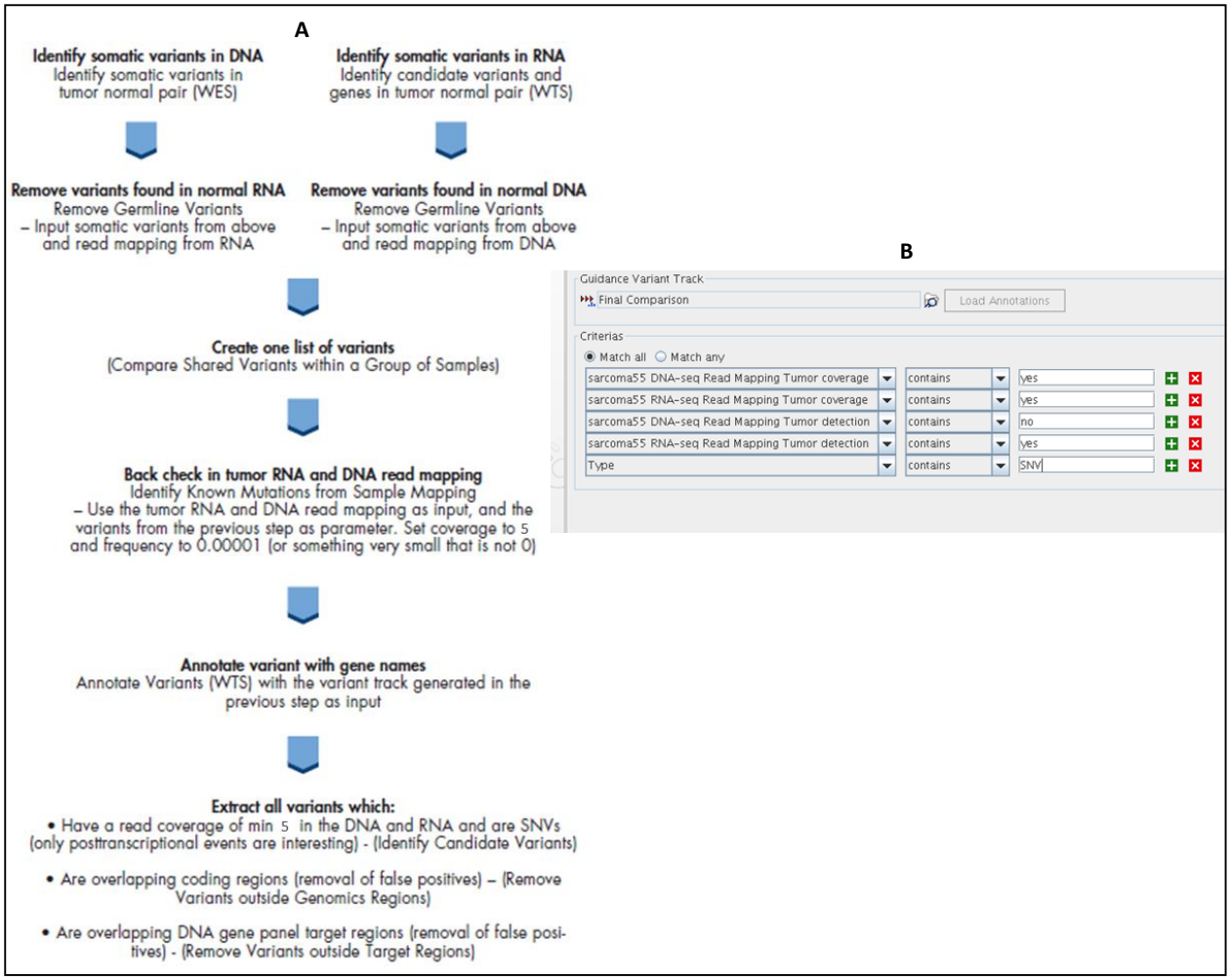


Figure 6.14. Steps to identify RNA editing variants specific to tumour RNAseq. A, *firstly, the somatic mutations were identified in the tumour DNA and RNA. Then the germline variants found in normal RNA were removed from the somatic DNA list, and the germline DNA variants were removed from the somatic RNA list. Create one list and annotate with gene names. Then extract the variants to choose the coverage and to remove false positives and variants outside target region.* ***B,*** *shows the variant track parameters. They should have coverage in DNA and RNA. They must be SNVs only and detected in the RNA but not DNA.*

The mutation types of these editing SNVs are different in the three patients. In patient 55, C>T is the highest edit, then G>A (C>T in the other strand). In patient, 66T>C is the highest edit type and then A>G. In patient 73, T>C is the edit and then C>T (figure 6.15). All of these types suggest that they may have resulted from the actions of *ADAR* and *APOBEC* enzymes.

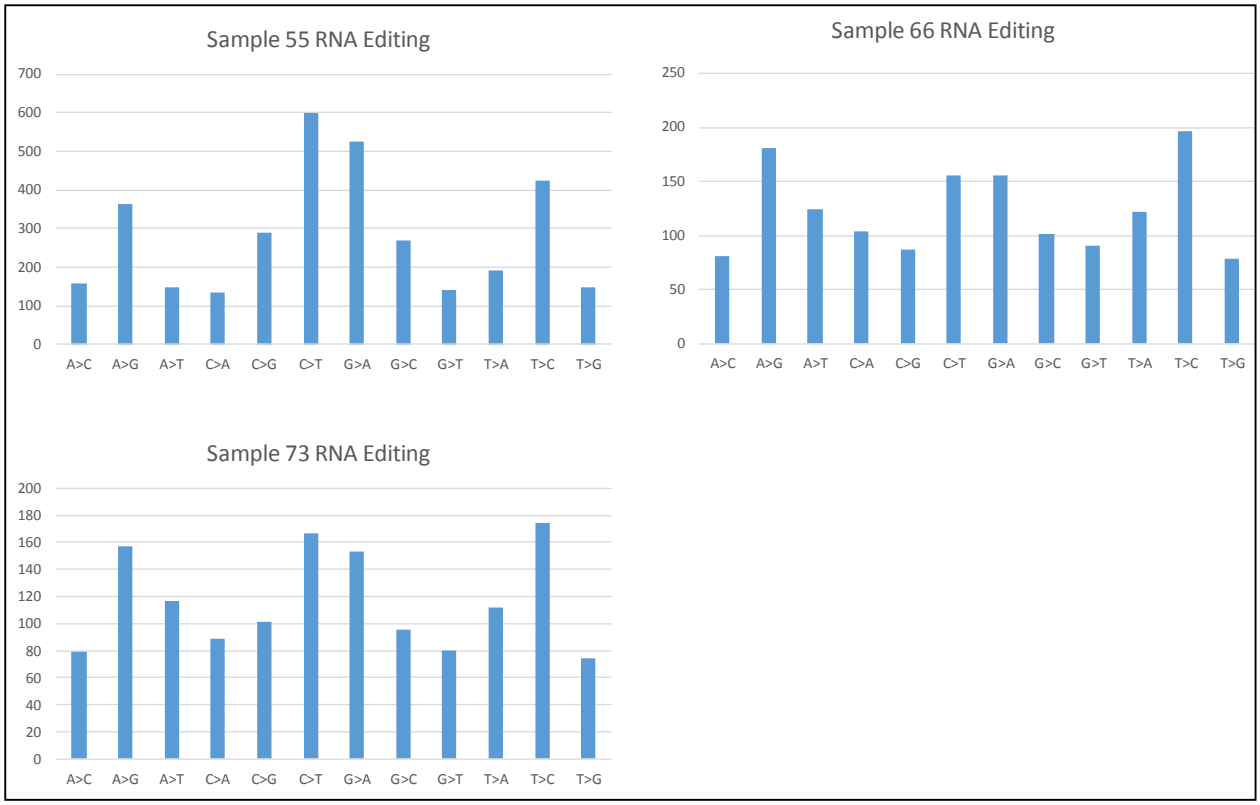


Figure 6.15. The mutation types of RNA editing SNVs in patients 55, 66 and 73. *The plot shows the number of each SNV type in each patient.*

6.2.4.8 Validation some of the RNA edits in patient 55 by deep sequencing

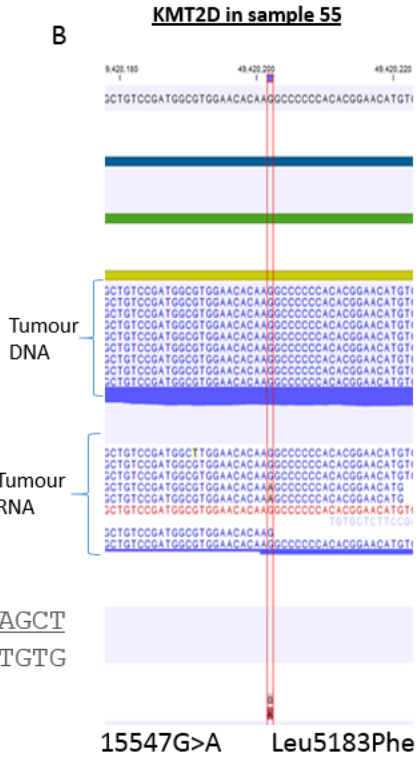
We chose 10 edits in patient 55 for further validation by deep sequencing (fig 6.16). We chose the primer sequencing around the edit site, and linked an adaptor sequence to these primers. We amplified the regions directly from the tumour RNA of patient 55 using the superscript one-step RT-PCR kit. We sent the PCR products to the laboratory in Brno to perform deep sequencing on the amplified regions.

We analysed the results of the deep sequences using the CLC software and found many variants almost at each base of the amplified regions, with different frequencies, including the base of edit. Thus, this method seems not to be the most suitable method to validate the RNA editing.

A

Chromosome	Region	Type	Reference	Allele	Tumour frequency	Gene name
6	33289002	SNV	A	G	33.33	DAXX
11	116633907	SNV	T	A	28.57	BUD13
16	30958168	SNV	T	C	25	FBXL19
21	30341916	SNV	T	A	25	LTN1
X	107084433	SNV	C	T	23.33	MID2
2	169622111	SNV	T	A	22.22	CERS6
15	41029488	SNV	T	A	22.22	RMDN3
18	8636316	SNV	C	A	22.22	RAB12
7	100730901	SNV	T	C	20	TRIM56
12	49420202	SNV	G	A	20	KMT2D

B



C

GCAGTGGCACTATGAAAGTCAGCCATCTGGTGAGGCAGCAGCT
GTCCGATGGCGTGGAACACAAGGCCCCCCACACGGAACATGTG
CAGCCGTTCTCCCCGCTGAATGATGCTAGCGATT

Figure 6.16. Choosing 10 RNA edits for validation. ***A,*** *table shows the 10 RNA edits chosen for validation by deep sequencing. It shows the chromosomal number and the SNV region, the frequency of the mutation in the RNA (count/coverage) and the gene name.* ***B,*** *the CLC Browser of the edit in the tumour RNA sequence of the KMT2D gene compared to the tumour DNA sequence and Hg19.* ***C,*** *the sequence of the region of KMT2D gene where the edit is (red G) and the primers chosen to amplify the region (underlined).*

6.2.4.9 RNA editing in *BLCAP* gene

We compared the genes with of non-synonymous edits in the three patients, and found 9 genes in common: *NBPF10, NBPF1, ASH1L, RUSC1, NBPF10, PLOD2, BLCAP, PSD3* and *MEGF8.*

We took BLCAP (bladder cancer-associated protein) as an example because this gene was reported to be a novel candidate tumour suppressor gene identified in human bladder carcinoma [125]. It was shown that BLCAP overexpression could inhibit cell growth by inducing apoptosis in HeLa cells, which suggests that the alteration in BLCAP may play an important role in tumourigenesis [125].

There were 6 non-synonymous mutations detected in the BLCAP gene in the three patients: one in patient 55, two in patient 66 and three in patient 73 (table 6.4). Tyr2Cys was detected in 66 and 73. Gln5Arg was detected in 55 and 73. Lys15Arg detected only in 66 and Pro19Thr only in 73. Three of these mutations: Tyr2Cys, Gln5Arg and Lys15Arg, have been detected

previously as RNA editing in BLCAP pre-mRNA amplified from human bladder [121]. The same study found that these editing events occurred within the most highly conserved region of the protein, within the N-terminus (figure 6.17), and were driven by the ADAR enzymes [121]. Pro19Thr edit in patient 73 has not been reported before, and is different from the other edits, as the previous edits were all A>G, but this is a C>A.

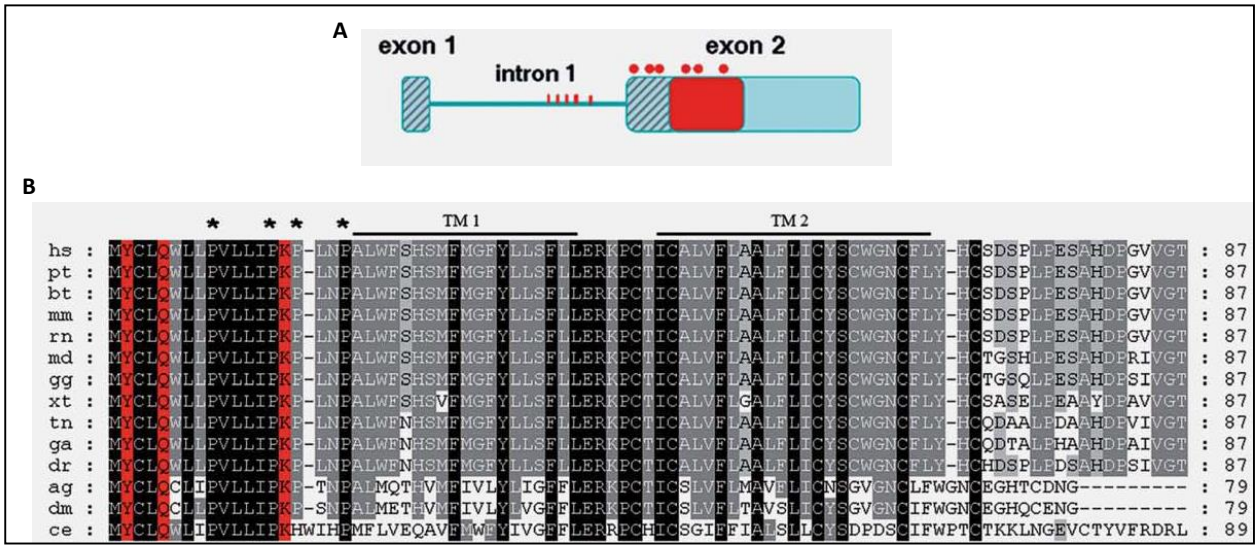


Figure 6.17. Structure of BLCAP gene and its conserved regions among species. A, *schematic representation of the BLCAP pre-mRNA. Exon 1 encoding the 5'UTR is represented in light blue with grey bars followed by the intron that divides the 5'UTR in 2. Exon 2 encodes the remaining 5'UTR (light blue with grey bars), the coding sequence (red rectangle) and the 3'UTR (light blue). Red dots indicate the editing sites in BLCAP mRNA (5'UTR and coding sequences). Red lines indicate the editing sites identified within the intron.* ***B****, multiple alignments of BC10 proteins. Initials correspond to hs, Homo sapiens; pt, Pan troglodytes; bt, Bos taurus; mm, Mus musculus; rn, Rattus norvegicus; md, Monodelphis domestica; gg, Gallus gallus; xt, Xenopus tropicalis; tn, Tetraodon nigroviridis; ga, Gasterosteus aculeatus; dr, Danio rerio; ag, Anopheles gambiae; dm, Drosophila melanogaster; ce, Caenorhabditis elegans. Identical amino acid conservation among species is indicated in black. Residues identical in most of the species analysed are indicated in grey. Amino acids found to be edited are Indicated in red. Black lines above the alignment indicate the hypothetical transmembrane domains (TM). Stars indicate the proline-rich motif at the N-terminus [121].*

Although some of the edits have been reported to occur in normal tissues, there may be differences in percentage of the editing events in the tumour and normal tissues due to different expression of editing enzymes in the two tissues. A study that integrated RNAseq and DNAseq data derived from two pairs of hepatocellular carcinoma (HCC) tissues, found that *BLCAP* is a novel editing gene and has over-editing expression in 40.1% HHC tissues, compared to adjacent liver tissues [126]. The same study has shown that the RNA-edited BLCAP, when compared to the wild type, stably promotes cell proliferation (including cell growth, colony formation in vitro, and tumoringenicity in vivo).

Sample	Mutation	Amino Acid change	Frequency
55	14A>G	Gln5Arg	7.14
66	5A>G	Tyr2Cys	18.75
	44A>G	Lys15Arg	10.71
73	5A>G	Tyr2Cys	35.29
	55C>A	Pro19Thr	11.11
	14A>G	Gln5Arg	11.11

Table 6.4. RNA editing SNVs in BLCAP gene in three UPS patients*. The table shows the type of mutations, which is A>G in all of them except one C>A in patient 73. The amino acid change and the frequency (count/coverage) of the mutation in the tumour RNA sequence are shown.*

Sample	Chromosome	Region	Type	Count	Coverage	Frequency %	Control coverage	Mutation	Amino acid change	Gene name
55	1	16893721	SNV	4	16	25	5	T>G	Phe931Cys	*NBPF1*
55	1	155979344	SNV	16	44	36.36	16	70C>T	Thr180Met	*SSR2*
55	1	1249715	SNV	52	57	91.23	25	226G>A	Ala144Thr	*CPSF3L*
55	1	113232660	Deletion	65	74	87.84	32	Del C	Arg260fs	*MOV10*
55	1	32209833	SNV	62	71	87.32	39	C>T	Arg350Trp	*BAI2*
55	1	167666379	SNV	88	104	84.62	47	518G>A	Arg173Gln	*RCSD1*
55	1	114453966	SNV	168	191	87.96	61	C>T	Thr251Met	*DCLRE1B*
55	1	42050112	SNV	191	220	86.82	73	357G>A	Trp119*	*HIVEP3*
55	3	49044826..49044828	Deletion	81	91	89.01	20	del CTT	Leu19del	*WDR6*
55	3	16368293	SNV	116	123	94.31	24	C>T	Val377Ile	*RFTN1*
55	3	36940696	SNV	72	164	43.9	40	205G>A	Ala69Thr	*TRANK1*
55	3	48474142	SNV	298	313	95.21	77	585A>T	Lys195Asn	*CCDC51*
55	3	184910772	SNV	62	250	24.8	101	1126G>A	Val376Ile	*EHHADH*
55	4	54319248..54319249	Deletion	6	97	6.19	24	del AG	Arg481fs	*FIP1L1*
55	4	73156712	SNV	96	105	91.43	38	G>C	Val931Leu	*ADAMTS3*
55	4	16035036	SNV	99	113	87.61	55	C>T	Arg134Cys	*PROM1*
55	5	115822460	SNV	55	60	91.67	14	C>A	Ser316Tyr	*SEMA6A*
55	5	180661266	SNV	21	54	38.89	27	G>T	Val462Phe	*TRIM41*
55	5	176083740	SNV	19	67	28.36	40	G>A	Val116Met	*TSPAN17*
55	5	306800	SNV	76	88	86.36	42	G>A	Val98Ile	*PDCD6*
55	5	88057085	SNV	47	150	31.33	63	G>A	Val107Ile	*MEF2C*
55	6	32609207	SNV	8	16	50	28	C>T	Ala68Val	*HLA-DQA1*
55	6	32609236..32609237	Deletion	11	20	55	29	Del GG	Gly78fs	*HLA-DQA1*
55	6	32551935..32551937	MNV	15	107	14.02	29	Del TAC ins GTG	Tyr107Val	*HLA-DRB1*
55	6	32609239	Deletion	12	20	60	30	Del G	Gly79fs	*HLA-DQA1*
55	6	32609216	SNV	8	16	50	30	EG>T	Trp71Leu	*HLA-DQA1*
55	6	32551948..32551949	Replacement	18	116	15.52	30	Del GC ins A	Ala103fs	*HLA-DRB1*

55	6	32609233	SNV	11	20	55	31	T>C	Phe77Leu	*HLA-DQA1*
55	6	32609212..32609213	MNV	8	16	50	31	Del CG ins AA	Arg70Lys	*HLA-DQA1*
55	6	32609221..32609222	MNV	8	18	44.44	31	Del GA ins CT	Glu73Leu	*HLA-DQA1*
55	6	32551955	Deletion	21	120	17.5	31	Del C	Arg101fs	*HLA-DRB1*
55	6	32609231	SNV	11	20	55	32	A>G	Lys76Arg	*HLA-DQA1*
55	6	32609227..32609228	MNV	10	19	52.63	32	Del AG ins CA	Ser75His	*HLA-DQA1*
55	6	35280196	SNV	30	83	36.14	33	G>A	Val89Ile	*DEF6*
55	6	32714125	SNV	5	100	5	33	A>G	Gln241Arg	*HLA-DQA2*
55	6	170871038..170871040	Deletion	86	88	97.73	38	Del CAA	Gln95del	*TBP*
55	6	56354402	SNV	102	118	86.44	38	T>A	His6601Gln	*DST*
55	7	138554437	SNV	7	36	19.44	8	G>A	Arg1541His	*KIAA1549*
55	7	44611098	SNV	68	141	48.23	28	C>T	Arg295Cys	*DDX56*
55	7	21939648	SNV	66	137	48.18	31	C>T	Arg4405Cys	*DNAH11*
55	7	107591763	SNV	110	235	46.81	49	C>T	Thr1124Met	*LAMB1*
55	8	101717237	SNV	14	41	34.15	7	G>A	Gly579Ser	*PABPC1*
55	8	145107704	SNV	74	110	67.27	21	C>T	Arg1040Cys	*OPLAH*
55	8	67428237	SNV	25	113	22.12	39	C>T	Arg184Trp	*C8orf46*
55	8	8234126	SNV	113	180	62.78	46	G>A	Arg598Gln	*SGK223*
55	8	28588774	SNV	90	132	68.18	48	G>A	Arg344Gln	*EXTL3*
55	9	34658519	SNV	143	158	90.51	55	C>T	Arg217Cys	*IL11RA*
55	9	107566931	SNV	121	134	90.3	65	C>T	Thr1512Met	*ABCA1*
55	10	126683243..126683244	MNV	11	94	11.7	33	Del GG ins CA	Gly192Gln	*CTBP2*
55	10	69571283	SNV	48	59	81.36	59	C>T	Thr99Met	*DNAJC12*
55	11	65408583	SNV	34	39	87.18	11	C>T	Thr64Met	*SIPA1*
55	11	118390715	SNV	16	44	36.36	14	C>T	Arg3789Cys	*KMT2A*
55	11	47652637	SNV	11	67	16.42	18	A>C	Ile135Leu	*MTCH2*
55	11	47652609	SNV	12	69	17.39	20	T>C	Ile153Thr	*MTCH2*
55	11	47652618	SNV	13	67	19.4	21	T>C	Val141Ala	*MTCH2*
55	11	47652591	SNV	10	67	14.93	22	A>T	Tyr159Phe	*MTCH2*
55	11	26463504	SNV	65	76	85.53	35	C>T	Ser29Leu	*ANO3*
55	11	115049385	SNV	34	98	34.69	36	C>T	Arg369Cys	*CADM1*

55	11	66456617	SNV	68	156	43.59	52	G>A	Arg1995His	*SPTBN2*
55	11	14899700	SNV	162	174	93.1	58	C>G	Gln259Glu	*CYP2R1*
55	12	53491580..53491581	MNV	24	27	88.89	11	Del GC ins AT	Ala27Met	*IGFBP6*
55	12	112605710	SNV	67	73	91.78	19	C>T	Arg3902Cys	*HECTD4*
55	12	112103530	SNV	75	83	90.36	49	G>A	Arg240His	*BRAP*
55	12	120934287..120934289	Deletion	155	182	85.16	55	Del GGT	Val22del	*DYNLL1*
55	14	106237703	SNV	5	32	15.62	5	G>C	Cys13Ser	*IGHG3*
55	14	106237697	SNV	5	34	14.71	6	G>A	Arg15Lys	*IGHG3*
55	14	80328191	SNV	105	120	87.5	65	C>T	Arg1024Trp	*NRXN3*
55	15	85632630	SNV	26	59	44.07	15	G>A	Glu233Lys	*PDE8A*
55	16	87902857	SNV	4	22	18.18	5	A>G	Ile58Val	*SLC7A5*
55	16	1291608	SNV	4	42	9.52	19	A>G	His74Arg	*TPSAB1*
55	16	67298342	SNV	70	81	86.42	24	C>T	Arg644Trp	*SLC9A5*
55	16	15703519	SNV	112	136	82.35	51	G>A	Arg1107His	*KIAA0430*
55	16	88793197	SNV	173	190	91.05	64	C>T	Arg1209Trp	*PIEZO1*
55	17	56056604^56056605	Insertion	31	34	91.18	7	Ins GCAGCA	Gln172_His173insGlnGln	*VEZF1*
55	17	63687	SNV	46	50	92	22	G>A	Ala273Thr	*RPH3AL*
55	17	7577538	SNV	72	75	96	23	G>A	Arg248Gln	*TP53*
55	17	18832245	SNV	62	69	89.86	27	G>A	Arg309Gln	*PRPSAP2*
55	17	8052583	SNV	34	107	31.78	46	G>A	Arg300His	*PER1*
55	17	72436616	SNV	125	139	89.93	53	G>A	Arg279His	*GPRC5C*
55	19	14552150	SNV	24	41	58.54	18	G>A	Val85Ile	*PKN1*
55	19	44676278	SNV	54	121	44.63	24	C>T	Thr18Met	*ZNF226*
55	19	58965090	SNV	33	108	30.56	41	G>A	Val8Met	*ZNF324B*
55	19	15281200	SNV	57	127	44.88	52	G>A	Val1686Met	*NOTCH3*
55	19	3619224	SNV	76	167	45.51	58	C>T	Arg301Trp	*CACTIN*
55	20	42159477	SNV	58	65	89.23	25	G>A	Arg321His	*L3MBTL1*
55	20	4776548	SNV	43	56	76.79	43	G>A	Arg67His	*RASSF2*
55	21	43274912	SNV	137	318	43.08	66	G>A	Gly101Arg	*PRDM15*
55	22	40804095	SNV	30	114	26.32	53	G>A	Val534Met	*SGSM3*

55	22	47059926	SNV	55	189	29.1	65	G>A	Arg210His	*GRAMD4*
55	X	153052558	SNV	109	114	95.61	27	G>A	Arg242His	*IDH3G*
66	1	175129946	Deletion	37	56	66.07	14	Del G	Lys69fs	*KIAA0040*
66	1	175129948..175129955	Deletion	37	66	56.06	15	Del CAAGAAGA	Asn65fs	*KIAA0040*
66	1	41288034	SNV	37	76	48.68	15	G>A	Ala364Thr	*KCNQ4*
66	1	16907946	SNV	2	35	5.71	16	G>A	Ala450Thr	*NBPF1*
66	1	77627307	SNV	14	46	30.43	28	G>A	Ser225Asn	*PIGK*
66	2	233712227..233712229	Deletion	19	23	82.61	6	Del ACA	Gln1237del	*GIGYF2*
66	2	241991199	SNV	89	163	54.6	62	G>A	Gly592Arg	*SNED1*
66	3	195509515	SNV	7	19	36.84	6	A>T	Asp2979Val	*MUC4*
66	4	190862169	SNV	26	140	18.57	39	C>T	Ala2Val	*FRG1*
66	4	190862165^190862166	Insertion	24	139	17.27	40	Ins CTTC	Met1?	*FRG1*
66	5	76357599	SNV	2	15	13.33	24	G>T	Gly639Val	*AGGF1*
66	5	137681174	SNV	119	247	48.18	55	G>A	Arg266His	*FAM53C*
66	6	42897358..42897360	Deletion	3	52	5.77	11	Del TGC	Leu25del	*CNPY3*
66	6	16327915^16327916	Insertion	52	116	44.83	28	Ins GCA	Gln208_His209insGln	*ATXN1*
66	9	33385602..33385603	MNV	8	47	17.02	7	Del GT ins AG	Gly131Glu	*AQP7*
66	9	33385600	SNV	7	46	15.22	7	C>T	Pro132Leu	*AQP7*
66	10	96282182	SNV	6	22	27.27	5	C>G	Leu658Val	*TBC1D12*
66	10	126691583	SNV	17	144	11.81	45	G>A	Val102Met	*CTBP2*
66	10	126691636	SNV	14	205	6.83	62	C>A	Thr84Asn	*CTBP2*
66	11	125359559	SNV	52	158	32.91	42	G>A	Glu39Lys	*FEZ1*
66	11	92624013	SNV	113	357	31.65	80	G>A	Glu4352Lys	*FAT3*
66	12	6128898	SNV	3	42	7.14	6	T>G	Val1229Gly	*VWF*
66	12	58240163	SNV	17	108	15.74	29	C>G	Thr19Ser	*CTDSP2*
66	12	58240184	SNV	7	103	6.8	30	G>T	Arg12Met	*CTDSP2*
66	12	7045924	SNV	46	173	26.59	76	G>T	Gln498His	*ATN1*
66	13	58299231	SNV	87	214	40.65	41	C>T	Gln1095*	*PCDH17*
66	13	43180756	SNV	85	208	40.87	77	A>G	Asn146Ser	*TNFSF11*
66	14	105411971	SNV	3	18	16.67	15	A>G	Ser3273Gly	*AHNAK2*
66	14	105415655	SNV	6	18	33.33	17	C>G	Leu2045Val	*AHNAK2*
66	14	105355902	SNV	190	247	76.92	67	G>C	Glu1194Gln	*CEP170B*
66	16	87902867	SNV	4	44	9.09	8	C>G	Asn54Lys	*SLC7A5*

66	16	69942738	SNV	21	113	18.58	22	G>A	Arg330Gln	*WWP2*
66	16	15122780	SNV	6	111	5.41	28	G>C	Gly417Ala	*PDXDC1*
66	17	45234702	SNV	9	36	25	11	A>G	Asn175Ser	*CDC27*
66	17	45234706	SNV	8	35	22.86	13	C>G	Gln174Glu	*CDC27*
66	17	45234720..45234721	MNV	8	31	25.81	15	Del AA ins TG	Lys169*	*CDC27*
66	19	7619899	SNV	23	94	24.47	21	T>A	Ser881Thr	*PNPLA6*
66	22	41322341	SNV	17	113	15.04	27	G>A	Asp453Asn	*XPNPEP3*
73	1	16918486	SNV	2	31	6.45	7	G>A	Glu11Lys	*NBPF1*
73	1	16918424	SNV	6	57	10.53	18	C>G	Asn31Ly	*NBPF1*
73	1	26608831	SNV	104	272	38.24	28	G>A	Gly475Ser	*UBXN11*
73	2	67631831	SNV	59	60	98.33	8	G>A	Glu673Lys	*ETAA1*
73	3	49395674..49395679	Deletion	9	9	100	6	Del GGCGGC	Ala12_Ala13del	*GPX1*
73	3	127806598	SNV	56	121	46.28	22	G>A	Arg357Gln	*RUVBL1*
73	3	47476554	SNV	25	107	23.36	30	C>A	Pro66Th	*SCAP*
73	3	75787663..75787665	MNV	28	405	6.91	66	Del TTG ins ACA	Ile363_Glu364delinsAsnLys	*ZNF717*
73	3	75787671..75787673	MNV	32	421	7.6	67	Del CAA ins TGC	Lys361Ala	*ZNF717*
73	4	78740203^78740204	Insertion	24	26	92.31	5	Ins GTGT	Ser5fs	*CNOT6L*
73	4	190862169	SNV	17	152	11.18	33	C>T	Ala2Val	*FRG1*
73	4	190862167^190862168	Insertion	17	151	11.26	35	Ins CGA	Met1_Ala2insArg	*FRG1*
73	4	190862165^190862166	Insertion	17	154	11.04	35	Ins CTTC	Met1?	*FRG1*
73	6	32497960..32497962	MNV	9	45	20	16	Del AAG ins GTT	Lys14Val	*HLA-DRB5*
73	6	29911970	SNV	7	57	12.28	21	G>A	Gly231Ser	*HLA-A*
73	6	32557478..32557479	MNV	7	72	9.72	23	Del CG ins TT	Ala14Val	*HLA-DRB1*
73	6	31238909..31238910	MNV	10	140	7.14	37	Del AC ins CG	Thr187Arg	*HLA-C*
73	6	33048569	SNV	9	175	5.14	42	119T>A	Phe74Tyr	*HLA-DPB1*
73	6	26250829	SNV	84	168	50	61	C>T	Ala2Val	*HIST1H3F*
73	9	15422903	SNV	61	84	72.62	45	26C>G	Pro9Arg	*SNAPC3*
73	10	126682443	SNV	4	45	8.89	12	C>A	His298Asn	*CTBP2*
73	10	19884350	SNV	42	58	72.41	45	C>T	Pro320Ser	*C10orf112*
73	11	47660334..47660335	MNV	13	164	7.93	39	Del TG ins CA	Gly66Arg	*MTCH2*

73	11	47660354	SNV	11	171	6.43	39	A>T	Gln59Leu	*MTCH2*
73	11	76506673..76506675	Deletion	9	157	5.73	41	Del CTG	Leu9del	*TSKU*
73	11	117163801	SNV	40	162	24.69	56	A>T	Asn226Ile	*BACE1*
73	11	92088133	SNV	70	260	26.92	61	G>A	Gly952Asp	*FAT3*
73	12	58217421..58217422	MNV	10	123	8.13	34	Del AC ins GT	Asp87Gly	*CTDSP2*
73	12	58217814..58217815	MNV	12	133	9.02	65	Del TT ins GC	Phe157Ala	*CTDSP2*
73	12	58217803..58217804	MNV	19	168	11.31	71	Del TT ins AC	Cys192Arg	*CTDSP2*
73	13	98641358	SNV	61	81	75.31	34	T>G	Leu117Arg	*IPO5*
73	14	105409196	SNV	14	129	10.85	16	G>A	Asp4198Asn	*AHNAK2*
73	15	23931508	SNV	25	67	37.31	14	A>G	Asp286Gly	*NDN*
73	15	75659903	SNV	57	152	37.5	43	C>A	Cys100*	*MAN2C1*
73	16	1291608	SNV	8	66	12.12	18	A>G	His136Arg	*TPSAB1*
73	16	67236149	SNV	83	213	38.97	84	C>T	Thr461Met	*ELMO3*
73	17	45235583	SNV	3	14	21.43	6	T>G	Leu155*	*CDC27*
73	17	45235598	SNV	3	19	15.79	7	C>G	Ser150Cys	*CDC27*
73	17	45235616	SNV	3	27	11.11	9	T>G	*75Glu	*CDC27*
73	17	45235620	SNV	3	29	10.34	9	A>C	Ser143Arg	*CDC27*
73	17	45235653	SNV	3	40	7.5	11	G>A	Ala132Thr	*CDC27*
73	17	45235635	SNV	4	33	12.12	12	T>G	Tyr138Asp	*CDC27*
73	17	72773474^72773475	Insertion	6	101	5.94	30	Ins G	Glu4fs	*TMEM104*
73	19	23542290..23542291	MNV	10	22	45.45	8	Del CT ins GC	Leu1164Ala	*ZNF91*
73	19	1799763^1799764	Insertion	3	27	11.11	8	Ins T	Glu526fs	*ATP8B3*
73	19	22270833	SNV	10	69	14.49	16	T>A	Phe94Tyr	ZNF257
73	19	22270839..22270841	MNV	10	71	14.08	18	Del AAA ins CTG	Gln96_Lys97delinsProGlu	ZNF257
73	20	34285616	SNV	35	102	34.31	29	G>T	Arg105Leu	NFS1
73	20	46279815..46279823	Deletion	54	186	29.03	34	Del GCAGCAGCA]	Gln1200_Gln1202del	NCOA3

Table 6.2. The expressed somatic mutations in the RNAseq of patients 55, 66 and 73. *The table shows the chromosomal number and region of each mutation, the mutation type, count, coverage and frequency in the tumour DNAseq, coverage in the normal DNAseq, mutation, amino acid change and the gene name.*

6.3 Conclusion

Cancer genome sequencing has revolutionized our understanding of the genetic basis of cancer, the classes of mutagenic events that drive cancer development, and the identification of genetic drivers. However, we do not yet know how this mutated code is translated into an expressed phenotype. Identifying the expressed cancer genome, using RNAseq and protein quantitation methods such as mass spectrometry, provides a more accurate view of the state of the cancer tissue at the time of presentation in the clinic. I was fortunate to have the opportunity to use a new software tool (in collaboration with CLCbio) that focused on the current state of the art in RNAseq mutation software that exploited cancer DNA sequencing. This enabled us to identify the dominant RNAs that were cancer genome encoded, and led us to highlight ~15% of mutated cancer genes are expressed at the time of surgery. I suggest that these mutated pathways or genes represent more realistic targets for therapeutics or diagnostics than conventional mutated cancer genes. For example, a mutant gene might have been required very early in the evolution of the cancer and might not even be expressed at the time of surgery. In addition to the expressed cancer-encoded genome, we also were able to use the novel software to identify mutations in RNA that are not genome encoded. This is suggestive of RNA editing. Some of these genes were validated by Sanger sequencing and by examining RNAseq in cell lines. For example, we find that MAP4K5 is indeed edited in cell lines, thus providing a model gene to dissect stages in RNA editing in the future. Currently, we do not know how many of the cancer genes that are mutated, the expressed RNA that is genome encoded or the RNA that is edited, are expressed at the protein level. I have shown one example where the mutated *PDCD6* gene had detectable mutated RNA and mutated peptide using mass spectrometry. The full exploitation of mass spectrometry was beyond the scope of my PhD thesis, which was focused largely on DNA and RNA sequencing. It will nevertheless be interesting in future to determine how much of the potential mutation, whether coded from genome mutation or RNA editing, is translated into mutated protein. My current thesis data shows, using novel software designed for biologists and clinicians, that is it realistic to routinely define this expressed genome at the nucleic acid level to support future proteomics studies.

CHAPTER SEVEN
Studying the effects of the expression of wild-type and mutant ELMO1 in oesophageal adenocarcinoma cell lines

7.1 Abstract

Genomics and proteomics provide us with the tools with which to identify novel mutated genes and their associated proteins. The development of protocols to capture new signalling pathways driven by mutated genes and proteins will be valuable for exploiting genomic information. In this chapter, methods with which to explore functions of a novel mutated protein identified from genomic studies are set up. I would have aimed to develop this with mutations from UPS, but do to time constraints of the PhD training, I in parallel developed this method using mutations generated in OAC at the start of my PhD training. The mutated protein (ELMO1) was identified and chosen from OAC genomics, and the one chosen is a gain-of-function oncogenic mutation. The mechanism whereby it operates as a gain-of-function mutant is not known and the interactomic methodologies set up in this chapter aim towards finding this out. OAC is characterized by an early invasion pathway that is linked to high frequency of gain-of-function mutations in the RAC signalling pathway that mediates metastasis. One of the key high frequency genetic mutations is in the DOCK–ELMO pathway, and the mechanisms describing its gain-of-function signalling are not well defined. We developed an SBP-tagged affinity purification method, in combination with label-free SWATH-MS, to identify novel binding proteins of the gain-of-function mutant protein ELMO1, the product of a key metastatic gene. This method identified an elevated interaction with other oncogenic proteins encoded by the AGR2 and DCD genes and validates this proteomics discovery platform for further advancing the discovery of functions of new mutated proteins.

7.2 Introduction

7.2.1 mutant genes in OAC

OAC has undergone a significant and rapid increase in incidence in the Western world. The reasons behind this increase are not completely understood. Because the disease is usually at an advanced stage at the time of diagnosis, the prognosis of OAC remains poor, with an overall five-year survival of ~18% [127]. In 2013, a study analysed the mutation spectra from whole-exome sequencing of 149 OAC tumour–normal pairs, 15 of which were also subjected to whole-genome sequencing, a step expected to lead to identification of new therapeutic target for OAC [41]. By using MuTect and Indelocator to identify somatic mutations, the study found 26 genes to be significantly mutated, with two known tumour suppressors in OAC: *TP53* and *CDKN2A* being the most significant (fig 7.1). With the exception of *ARID1A, PIK3CA* and *SMAD4*, no other significantly mutated gene had previously been implicated in OAC, although several had been implicated in other cancers [41].

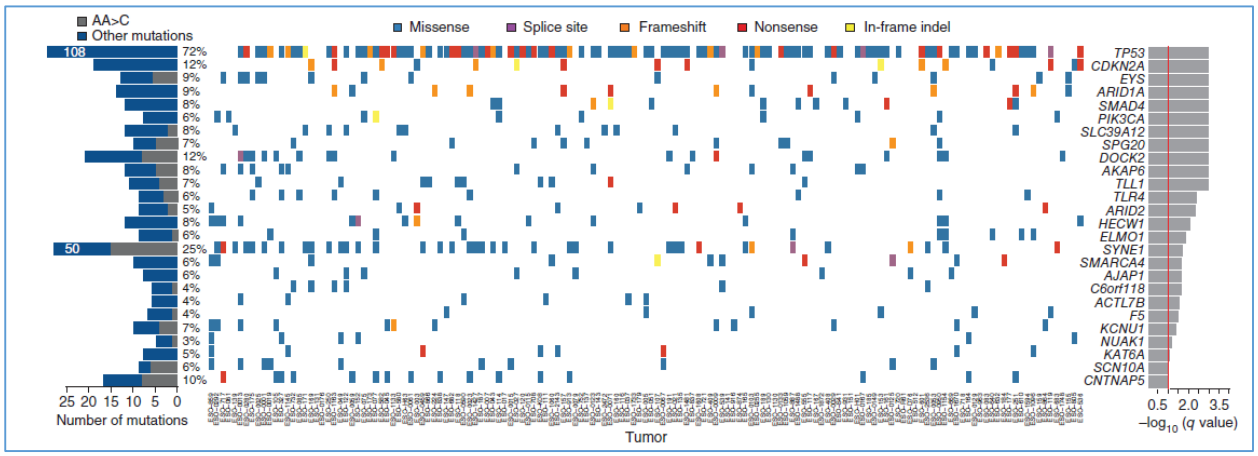


Figure 7.1. Significantly mutated genes in OAC as identified by whole-exome sequencing*. Mutations in significantly mutated genes coloured by the type of coding mutation. Each column denotes an individual tumour, and each row represents a gene. Left, number and percentage of samples with mutations in a given gene. The grey bar represents the number of transversions at AA sites in a gene. Listed numbers within bars represent values exceeding the scale. Right, $-\log_{10}$(q values) for the significance level of mutated genes shown for all genes with FDR q < 0.1 [41].*

7.2.2 Mutations in *ELMO1* and *DOCK2* in OAC samples

Two of the significantly mutated genes are engulfment and cell motility 1 (*ELMO1*) and dedicator of cytokinesis 2 (*DOCK2*), which encode dimerization partners and intracellular mediators of the Rho family GTPase, RAC1. Although no RAC1 mutations were identified, either ELMO1 or DOCK2 was mutated in 25 OAC samples (17%), with two samples having mutations in both genes and two samples having two independent mutations in DOCK2 (fig 7.2a, b). A single amino acid, Lys312 of ELMO1, was affected by mutation in three tumours (fig 7.2a), which suggests a gain-of-function mutation. Because aberrant RAC1 activation has been implicated in malignant transformation in other cancer types, mainly by enhancing cellular motility, recurrent mutations in these genes may be functionally important [41].

7.2.3 Gain of function mutations in ELMO1

The mutations in ELMO1 were examined. Wild type (wt) and mutant ELMO1 constructs were generated and introduced into NIH/3T3 cells (mouse embryonic fibroblast cells). Compared to a green fluorescent protein (GFP) control, wt ELMO1 expression increased invasion by sevenfold (fig 7.2c). ELMO1 alterations (p.Phe59Leu, p.Lys312Glu, p.Lys312Thr, p.Lys349Arg and p.Thr421Asn) resulted in further significant increases (two to sevenfold) in invasion compared to wt ELMO1 (fig 7.2c). These results suggest that *ELMO1* mutations can increase invasiveness and potentially contribute to tumorigenesis in OAC [41].

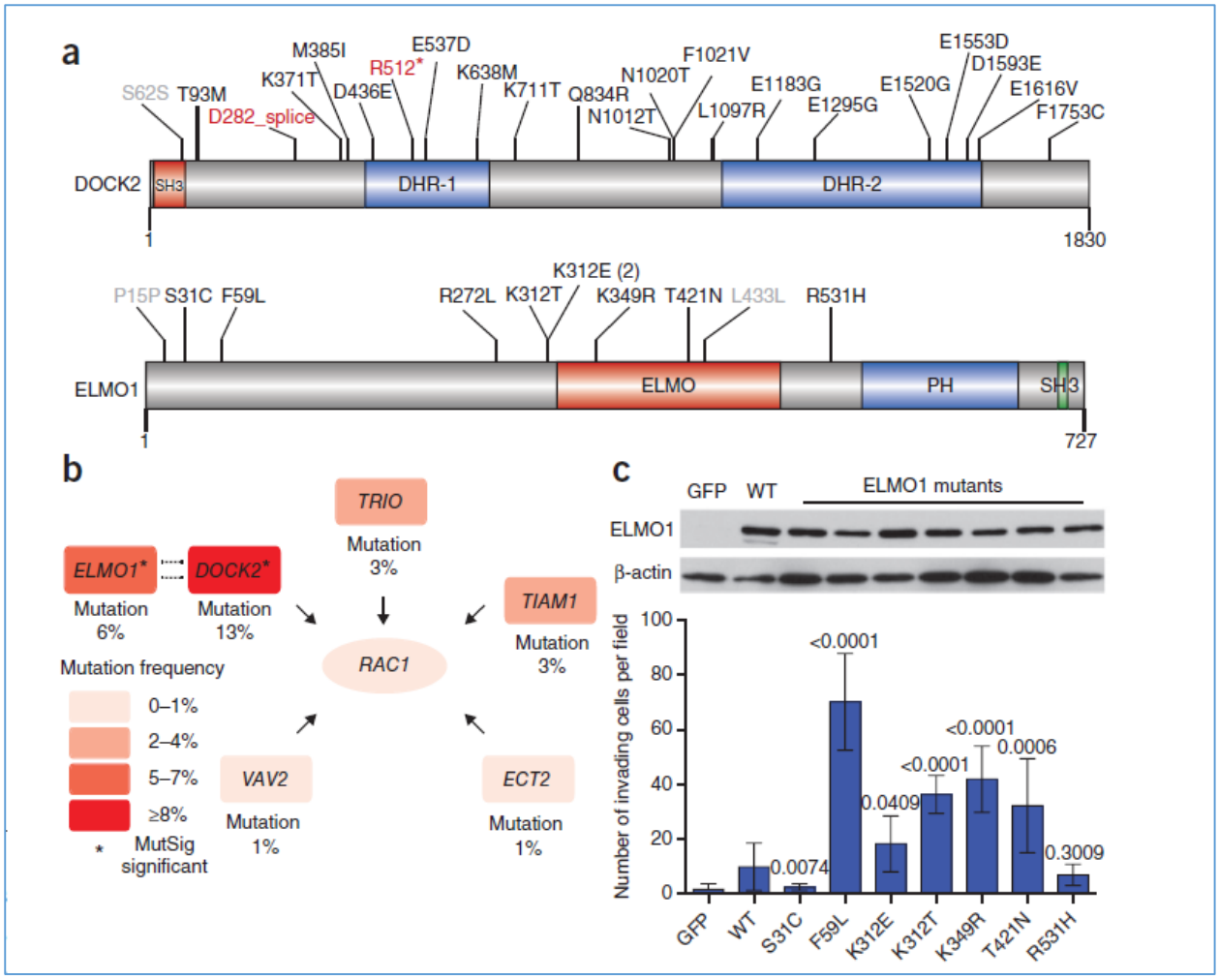


Figure 7.2. Recurrent somatic alterations in ELMO1, DOCK2 and other RAC1 GEFs. ***A,*** *schematics of protein alterations in DOCK2 and ELMO1 detected by whole-exome sequencing. Coding alterations in OAC are coloured either in black (missense) or red (splice site or nonsense); silent mutations are depicted in grey. Conserved domain mapping is from UniProt; SH3, SRC homology 3; DHR, Dlg homologous region, ELMO, engulfment and cell motility; PH, Pleckstrin homology.* ***B,*** *sample mutational frequency of candidate ELMO1 and DOCK2 as well as other RAC1-activating GEFs in 145 whole-exome sequenced OACs.* ***C,*** *wild-type or mutant ELMO1 proteins (or GFP control) were expressed in NIH/3T3 cells using retroviral transduction with the pBabe vector. Protein expression was confirmed by immunoblot analysis. Cells were plated in Matrigel invasion chambers with medium containing full serum in the lower chamber only, and invading cells from four fields were counted. The numbers of invading cells from three independent replicates are shown. Error bars, s.d. P values compare mutant ELMO1 to wild-type protein, Student's t test [41].*

7.2.4 RAC activation by ELMO1-DOCK complex

DOCK proteins are guanine nucleotide exchange factors (GEFs) controlling the activity of RAC1/CDC42 during migration, phagocytosis, and myoblast fusion. There are 11 mammalian members of the DOCK family: DOCK180 and DOCK2-11. They share the DOCK-homology regions (DHR1 and DHR2), which are known as Docker domains and lack the DbI homology (DH) and pleckstrin homology (PH) domains typically present in the mammalian Rho-family GEFs [128]. ELMO1 (75 KDa) is a cytoplasmic adaptor protein that physically associates with members of the DOCK family of RAC-GEFs, of which DOCK180 and DOCK2 are the best characterized. ELMO1 contains a PH domain for interaction with DOCK proteins [129]. Several studies have shown that ELMO1 binding enhances DOCK signalling by increasing its RAC-GEF activity, membrane localization, and protein stability [130]. It was reported that ELMO1 modulated the RAC1 activation with Dock180 by at least three explicit mechanisms namely: helping DOCK180 stabilize RAC1 in its nucleotide-free transition state; protecting Dock180 from ubiquitylation; and targeting Dock180 to the plasma membrane to get access to Rac1[131]. A study has developed a Transwell migration assay with LR73 cells (ovary cell line) to investigate the role of ELMO1 alone, DOCK180 alone, and co-expression of both genes, on cell migration [132]. The study found that co-expression of DOCK180 and ELMO1 strongly promoted migration compared with control cells, whereas expression of DOCK180 alone or ELMO1 alone did not promote migration, indicating a requirement for both proteins for this effect (fig 6.3). To prove this migration was RAC dependent, they co-transfected ELMO1 and DOCK180 with a dominant negative form of RAC (RacT17N), and found inhibition of migration increased (fig 7.3), which suggests that the enhancement of migration with DOCK180–ELMO1 depends on RAC activation.

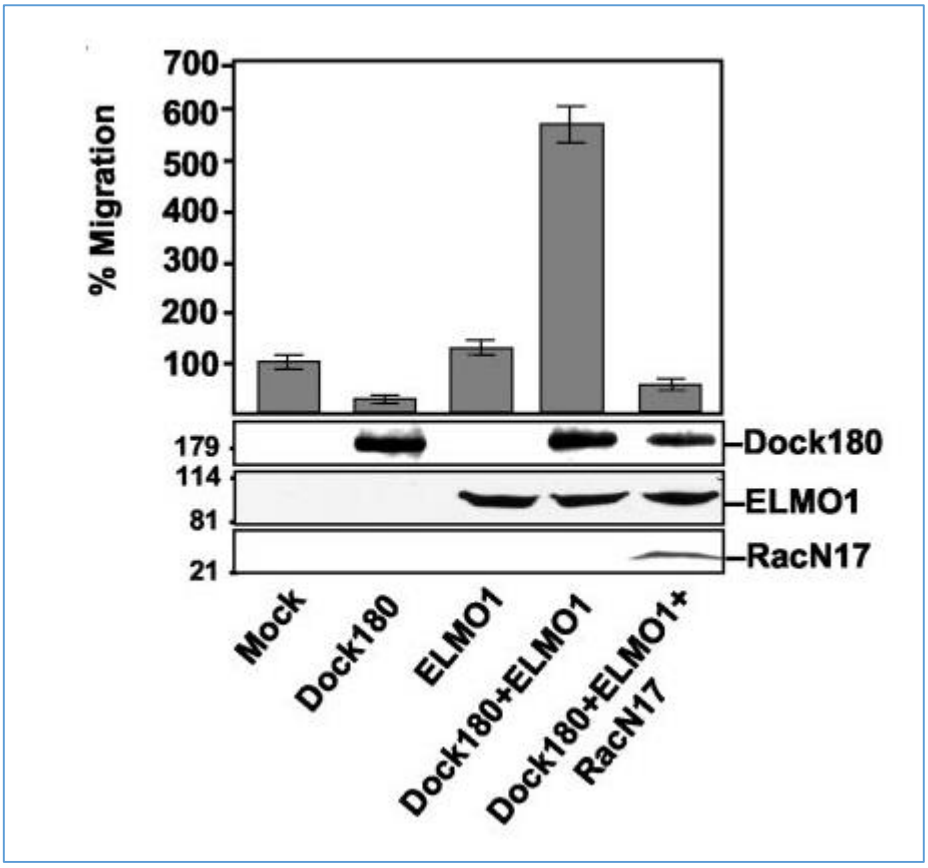


Figure 7.3. Dock180 and ELMO1 cooperate to promote RAC-dependent cell migration. *LR73 cells were transiently transfected with the indicated plasmids plus a luciferase reporter construct. After 24 h, 1–105 cells were plated in duplicate on top of a 24-well Transwell chamber filter and allowed to migrate for 6 h. Equal numbers of cells were also plated in a separate chamber without a filter to estimate total luciferase activity for each condition. The percentage migration was calculated by dividing counts in the bottom chamber (migrated cells) by the total cell counts for each condition. Each point represents the mean percentage ± S.E. of two duplicate migration chambers. The luciferase alone control is set at 100%. Aliquots of cells from each transfection condition were lysed and immunoblotted with anti-FLAG to confirm expression of Dock180 and RacT17N and anti-GFP to confirm expression of ELMO1 [132].*

7.2.5 The role of RAC1 in cellular processes

RAC1 is a member of the Rho GTPase family, which includes Rho, RAC1 and CDC42. These proteins classically regulate the machinery that controls the assembly and disassembly of cytoskeletal elements [133]. As a modulator of the cytoskeleton, RAC1 activity is critical for a number of normal cellular activities, including phagocytosis, mesenchymal-like migration and axonal growth. RAC1 also plays a major role in the moderation of other signalling pathways involved in cellular growth and cell cycle regulation. These RAC1-mediated activities appear central to the processes that underlie malignant transformation, including tumorigenesis, invasion, and metastasis [133]. The activation of RAC1 is through a GDP/GTP exchange mechanism catalysed by the GEFs resulting in an active, GTP-bound state [134]. The Rho GTPase GEFs are a large family of proteins that contain either a DH domain involved in nucleotide exchange or DHR domain that facilitates GEF function [134].

7.2.6 Structure of the ELMO1–DOCK complex

7.2.6.1 ELMO1

ELMO1 was previously identified as a mammalian orthologue of *Caenorhabditis elegans* CED-12. It is functionally divided into N-terminal and C-terminal parts. The N-terminal region of ELMO1 is necessary for membrane targeting. The C-terminal region of ELMO1 interacts with proteins such as DOCK2 and is required for modulating downstream molecules [135]. Although ELMO1 has no intrinsic catalytic activity, it mediates several cellular functions by providing a scaffold for signalling proteins or by regulating the activity of other proteins via protein–protein interactions [135]. Among vertebrates, ELMO1 is an evolutionarily highly conserved protein that contains several armadillo repeats, a putative pleckstrin homology domain, a DOCK180-binding motif and a proline rich region (fig 7.4) [129].

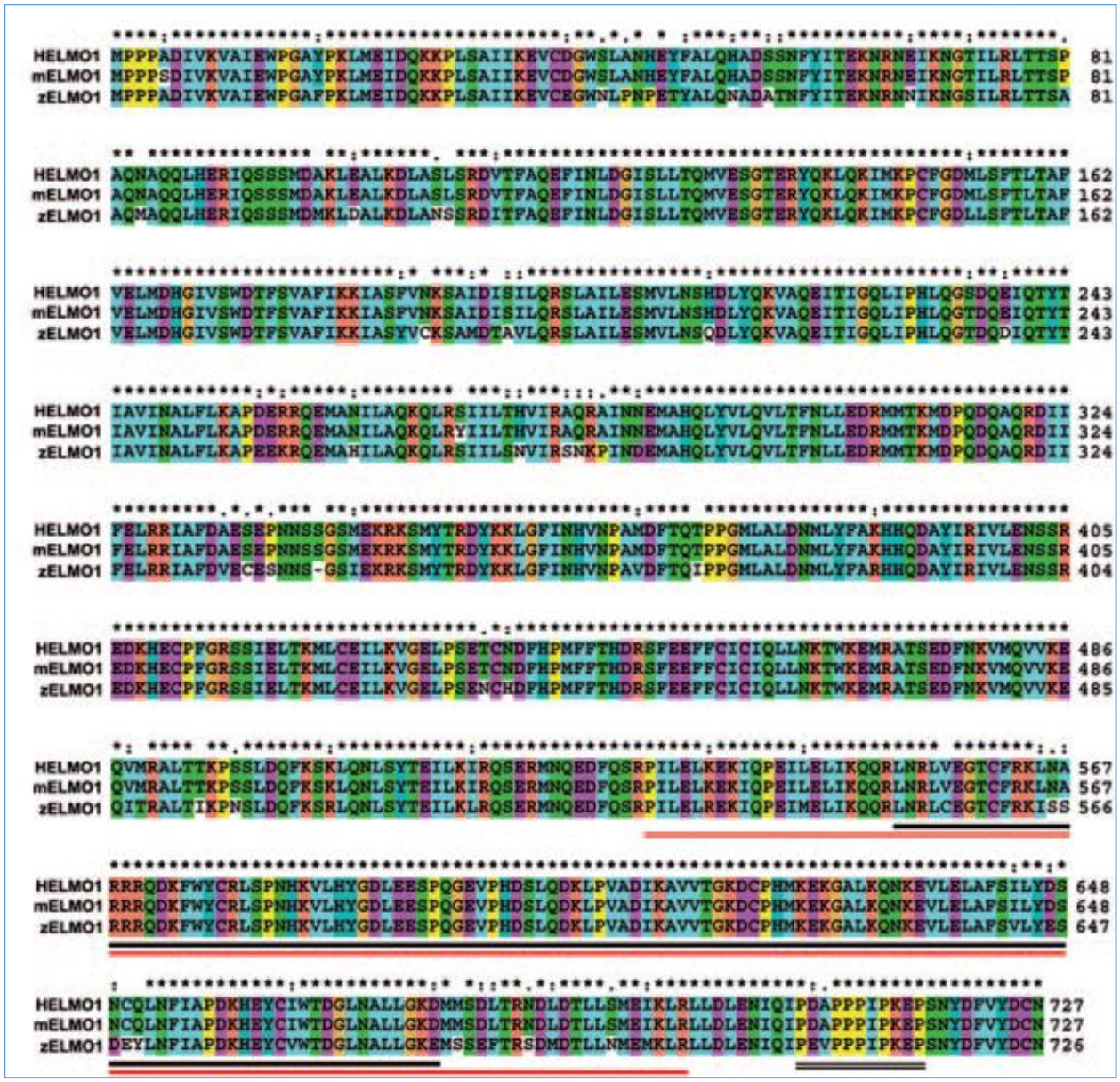


Figure 7.4. ELMO1 is a highly evolutionarily conserved protein. *Multiple amino acid alignment of human (HELMO1), mouse (mELMO1), and zebrafish (zELMO1) ELMO1. Numbers indicate amino acid positions. Indicated are the predicted pleckstrin homology domain (black line), DOCK180 binding domain (red line), and the proline-rich motif (double black line) [129].*

7.2.6.2 binding regions of ELMO1 and DOCK2

A study has identified the N-terminal 177-residue fragment and the C-terminal 196-residue fragment of human DOCK2 and ELMO1, respectively, as the mutual binding regions, and solved the structure of their complex at 2.1-A resolution (fig 7.5) [128]. The C-terminal Pro-rich tail of ELMO1 winds around the Src-homology 3 domain of DOCK2, and an intermolecular five-helix bundle is formed. Overall, the entire regions of both DOCK2 and ELMO1 assemble to create a rigid structure, which is required for the DOCK2-ELMO1 binding.

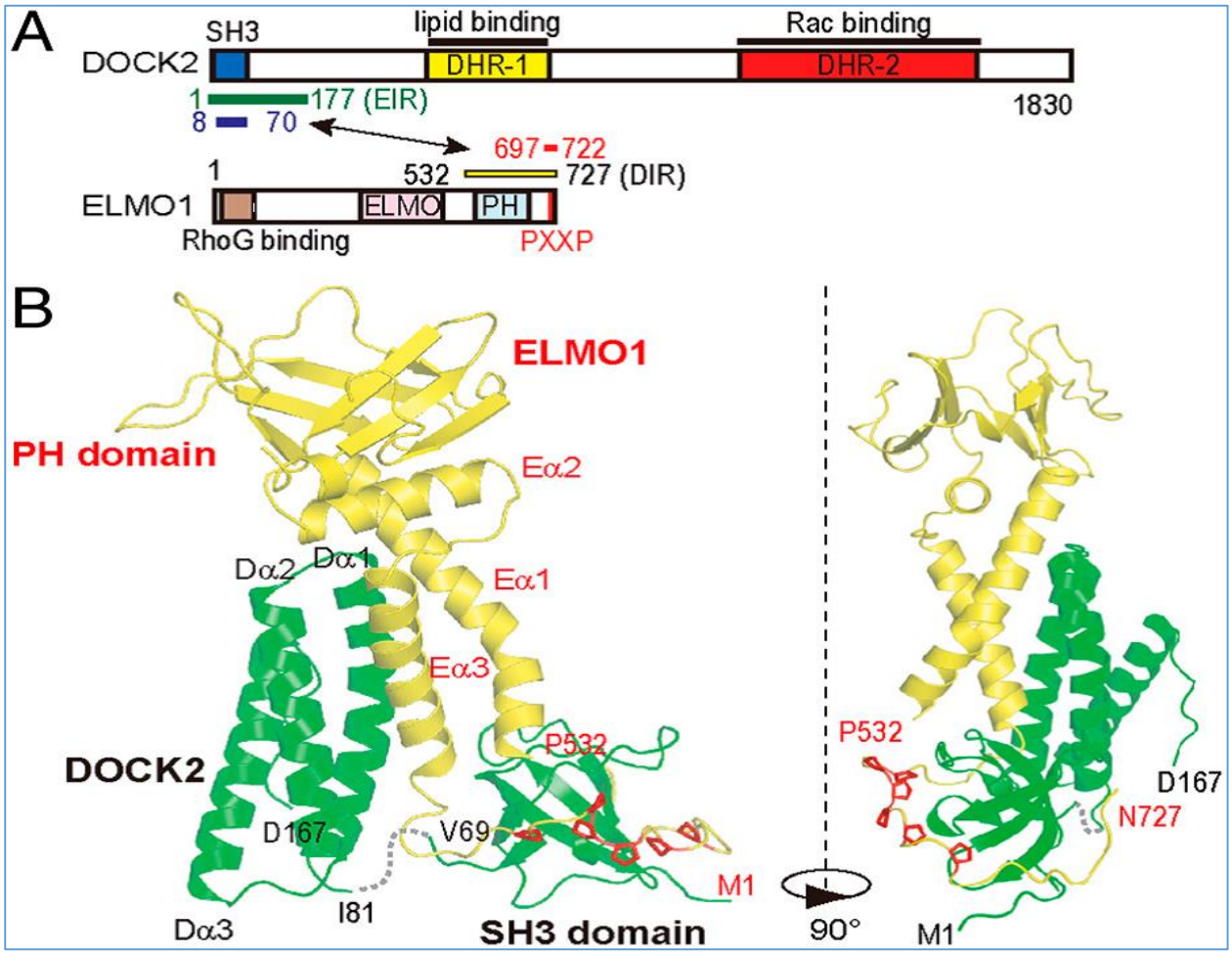


Figure 7.5. Mutually interactive regions of DOCK2 and ELMO1. ***A,*** *domain organizations of DOCK2 and ELMO1. Known protein interaction and functional regions are indicated. Blue and red bars indicate the DOCK2 and ELMO1 regions included in the NMR construct. Green and yellow bars indicate the regions used for the crystal structure determination. ELMO1-interacting region (EIR) and DOCK2-interacting region (DIR) indicate the ELMO1-interacting region and the DOCK2-interacting region, respectively.* ***B,*** *overviews of the crystal structure of the DOCK2–ELMO1 complex. Ribbon representations of the DOCK2–ELMO1 complex. DOCK2 is coloured green and ELMO1 is yellow. The six proline residues in the ELMO1 Pro-rich tail are coloured red. The two views are related by a 90° rotation about the vertical axis [128].*

7.2.6.3 An autoinhibitory mechanism in ELMO1 and DOCK2

Both ELMO1 and DOCK2 proteins were shown to have autoinhibitory mechanisms. In ELMO1, the ELMO1 inhibitory domain (EID) binds to the ELMO1 autoregulatory domain (EAD) resulting in an autoinhibited form of the protein (fig 7.6), and in DOCK2 the SH3 domain interacts with catalytic DHR-2 domain, which may block RAC1 access to the DHR-2 domain and cause autoinhibition of DOCK2 (fig 7.6). Binding of ELMO1 to RhoG promotes relief of autoinhibition in ELMO1, and the binding of ELMO EAD domain to DOCK2 relieves DOCK2 from autoinhibition, as shown in figure 7.6 [128, 136]. Double mutations in the EAD domain of ELMO1, Met692 and Glu693 to Ala, have been reported to disrupt the interaction between the EID and EAD, which leads to increased activity of ELMO1 [128].

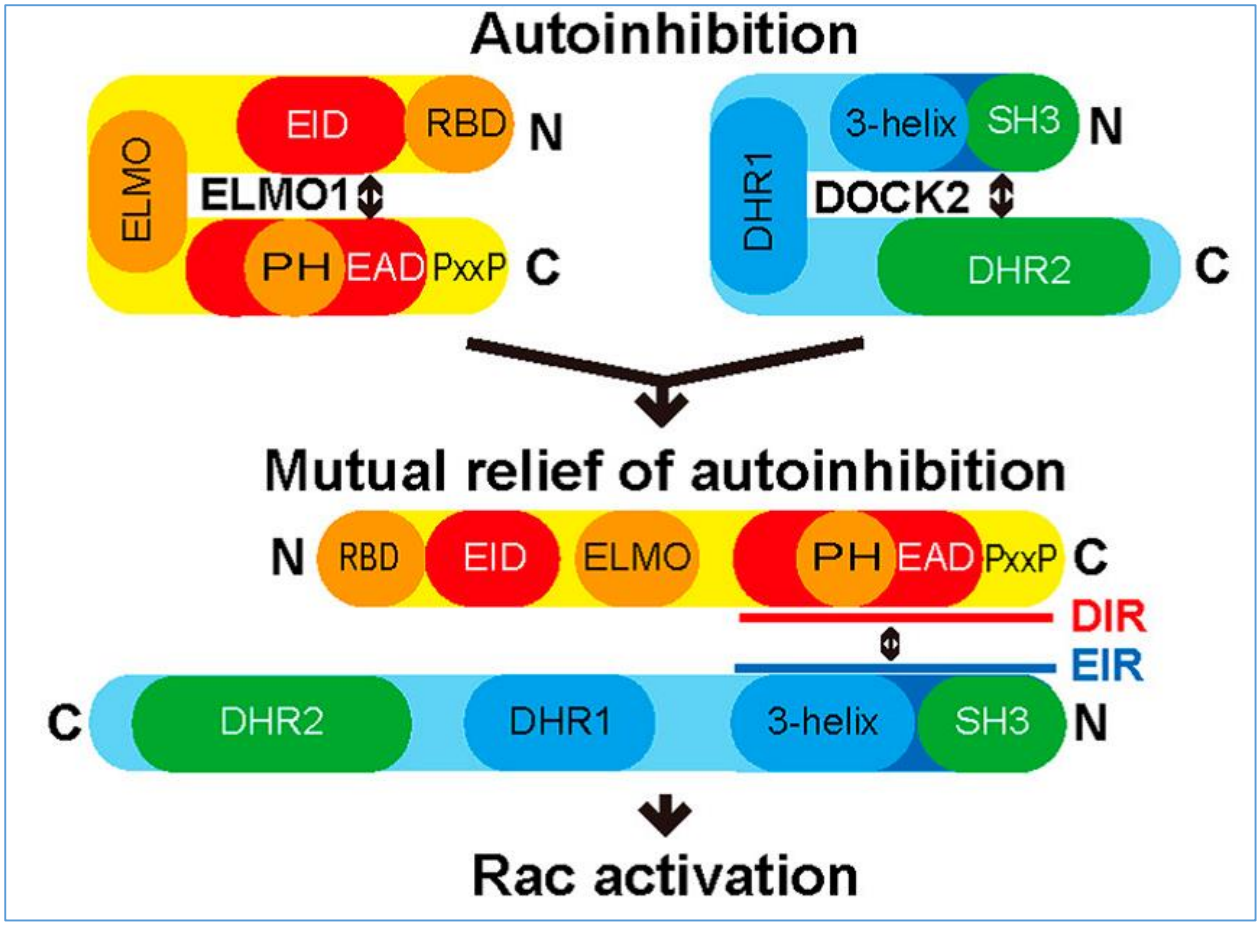


***Figure 7.6. Hypothetical model of the DOCK2–ELMO1 complex**. Schematic model of the mutual relief of DOCK2 and ELMO1 from their autoinhibited forms for RAC1 activation. EIR, ELMO1-interacting region; DIR, DOCK2-interacting region [128].*

7.2.7 ELMO1 binds to RhoG

In addition to the DOCK proteins, other proteins are reported to bind to ELMO1 such as RhoG. RhoG is a member of the Rho family of small GTPases and is involved in cellular morphological processes such as neurite outgrowth in neuronal cells, which also requires the activation of RAC1. The GTP-bound form of RhoG, but not the inactive form, specifically binds to the N-terminus of ELMO1. Thus ELMO1 is a mediator that links RhoG activation to RAC1 by interacting with both RhoG and DOCK [137, 138].

7.2.8 ELMO1 expression in cancer

A study of human glioma cells detected high level of ELMO1 and DOCK180 expression in actively infiltrating cells within the invasive regions compared with the central tumour areas of primary human glioma specimens. They also reported that the inhibition of endogenously expressed ELMO1 and DOCK180 attenuated the invasive behaviour of glioma cell lines concomitant with a reduction in the activated RAC1. Conversely, exogenous expression of ELMO1 and DOCK180 in low-level expressing glioma cells increased their capacity to migrate and invade, both in vitro and in the brain, emphasizing the importance of these molecules in the invasive process of these malignant cells [134].

Previous studies in Rhabdomyosarcoma (RMS) have shown that the more metastatic RMS subtype, which has a worse prognosis, also expresses higher levels of ELMO1. The down regulation of ELMO1 in these cells has been shown to decrease the invasive nature of the cells [139]. This indicates that ELMO1 may be involved in the migration and subsequent metastasis of RMS cells and may be considered as a novel target for treatment.

Another study aimed at identifying the role of ELMO1 in cell migration in ovarian cancer has reported that ELMO1 was mainly expressed in high-grade ovarian cancer tissues. Also, decreased colony formation and cell invasion were observed in ELMO1-RNAi cells, compared with the negative control, [131]. The study concluded that ELMO1 shows synergistic action in helping DOCK180 to activate RAC1 and promote cell motility, and thus promote untoward expansion and aggressiveness of ovarian cancer.

ELMO1 showed high expression in human breast cancer samples, and the expression correlates with lymph node metastasis and distant metastasis in breast cancer patient samples [140].

7.3 The aim and strategy of this study

The ELMO1 gene and its gain of function mutations detected in OAC samples may play a role in OAC metastasis and can be a therapeutic target for OAC patients. This therefore forms a "model" for developing interactomic methods to identify new functions for a mutated protein defined by genomics, which I would have aimed to do in UPS if training time permitted. The aim of this research was to study the effect of wt and mutant ELMO1 (F59L) proteins in OAC cell lines and to identify the mode of action of the gain-of-function mutation, but it can also apply to mutated genes in UPS, such as IL11R mutation or the PDCD6 mutation described in the previous chapters.

We tested the expression of ELMO1 in our key OAC cell lines: OE19, OE33 and FLO-1, and found that none of these cells express ELMO1. Therefore, we cloned ELMO1 cDNAs into human cell expression vectors and used them to develop artificial cell lines that express the wt and mutant ELMO1. We studied the effects of the wt and mutant ELMO1 on the growth of colonies in the transfected cell lines using a clonogenic assay. We have also developed an SBP-tagged affinity purification method, in combination with label-free SWATH MS, to identify novel binding proteins for the gain-of-function mutant ELMO1. This approach identified elevated interaction with other oncogenic proteins encoded by anterior gradient protein 2 (AGR2) and Dermcidin (DCD) genes. We further validated the interaction between ELMO1 and AGR2 proteins by a proximity ligation assay (PLA).

7.4 Results

7.4.1 Expression of ELMO1 in OAC cell lines

We grew OAC cell lines OE19, OE33 and FLO-1 in specific media in a 37°C incubator. When they were approximately 100% confluent we lysed them in urea lysis buffer. We measured the protein concentrations in each lysate by Bradford assay and ran 20 µg of protein on sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS–PAGE) for Western blot analysis. Jurkat cell whole-cell lysate (sc-2204 Santa Cruz Biotechnology) was used as a positive control for ELMO1 expression because it has been reported that this cell line expresses ELMO1 and DOCK2 [141]. By using an ELMO1 specific antibody (Abcam ab2239 goat polyclonal) we could not detect ELMO1 expression in any of three OAC cell lines, whereas there is expression of ELMO1 in Jurkat cells (fig 7.7). This result was supported by our failure to amplify ELMO1 by PCR from the cDNA made from the whole RNA purified from each of the OAC cell lines (data not shown). This means that ELMO1 is not expressed in these three OAC cell lines. The expression of ELMO1 may be in different locations in the tumour, compared with the OAC cell lines examined. Indeed, it has been reported that immunohistochemistry (IHC) analysis of primary human glioma specimens showed high levels of ELMO1 in actively invading tumour cells in the invasive area, but not in the central regions of these tumours [134].

To study the effects of *ELMO1* in the OAC cell lines we cloned *ELMO1* into the human expression vectors: pEGFP-C1 and pEXPR-IBA105.

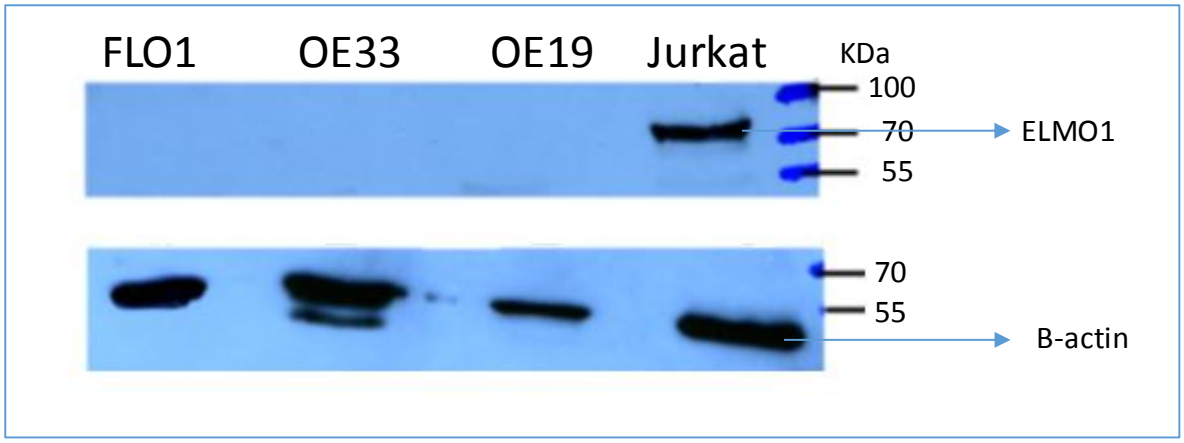


Figure 7.7. Western blot results of immunoblotting lysates of OAC cells and Jurkat cells with ELMO1 and β-actin antibodies. *OAC cell lines: OE19, OE33 and FLO-1 were lysed using urea lysis buffer. OAC protein lysates and whole Jurkat lysate (20µg) were run on a 12% gradient gel then transferred onto 0.2 µM nitrocellulose membrane. The membrane was incubated with goat polyclonal ELMO1 primary antibody at a dilution of 1/1000 for one hour at room temperature (RT), and after washing it was incubated with rabbit anti-goat secondary antibody (1/1000). The membrane was also incubated with rabbit polyclonal to beta Actin (β-actin) as positive control because Actins are ubiquitously expressed in all eukaryotic cells. Antibody signal was detected using enhanced chemiluminescence (ECL).*

7.4.2 Cloning of ELMO1 in pEGFP-C1 and pEXPR-IBA105 vectors

The Elmo1 gene was provided by Dr. Tilo (Heriot-Watt University) as a cloned gene in the pCR®-Blunt II-TOPO® vector. Elmo1 was amplified by PCR using the primers: ACCGGAATTCTATGCCGCCACCCGCGGAC and CGGTGGATCCTTAGTTACAGTCATAGACGAA with *EcoR1* and *BamH1* restriction sites respectively to clone it in the pEGFP-C1 vector, and GGGGGAATTCGATGCCGCCACCCGCGGACATCGTC and CCCCGGATCCTCAGTTACAGTCATAGACGAAGTC with *EcoR1* and *BamH1* restriction sites respectively to clone it in pEXPR-IBA105 Twin-Strep-tag vector (fig 7.8).

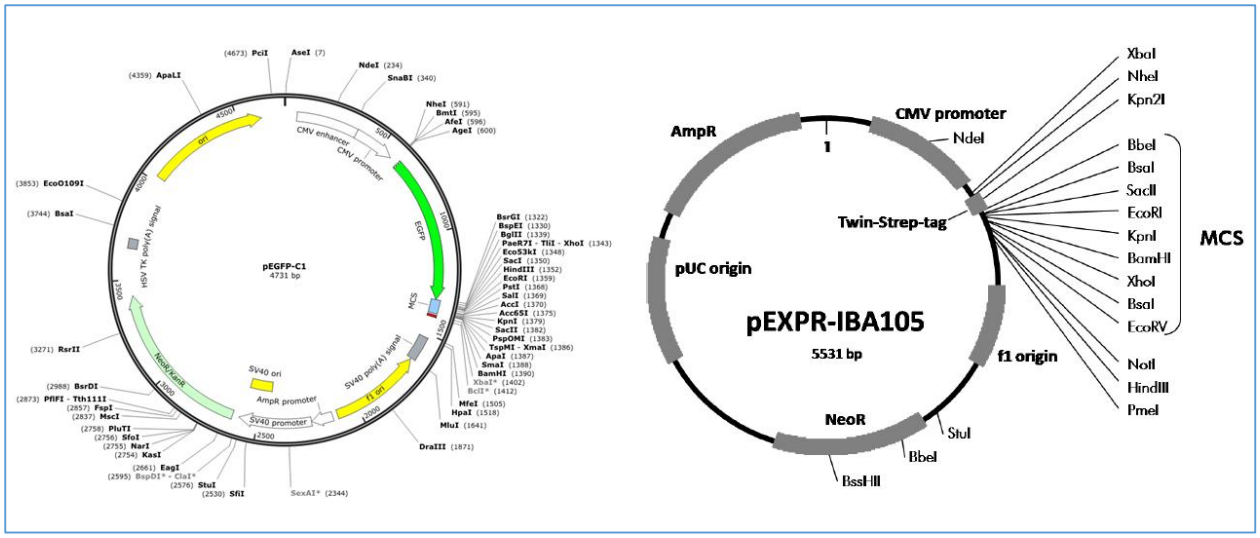


Figure 7.8. Diagram of pEGFP-C1 and pEXPR-IBA105 vectors in which ELMO1 was cloned using EcoR1 and BamH1 restriction sites. *Both vectors have neomycin resistance selectivity.*

The PCR product of Elmo1 and the two vectors (pEXPR-IBA105 and pEGFP-C1) were digested with *EcoR1* and *BamH1* restriction enzymes. After 3 h incubation at 37°C the digested DNA was run on a 1% agarose gel, specific bands were excised from the gel and DNA extraction was performed using a DNA extraction kit (Qiagen) followed by PCR purification.

A ligation reaction was carried out to clone Elmo1 into the two vectors.

DH5α cells were transformed with the ligation product and incubated at 37°C overnight. Colonies obtained were grown in LB media at 37°C overnight with shaking, and DNA mini-preps were prepared the following day. Mini-preps were digested with *EcoR1* and *BamH1* and analysed on a gel to determine which clone contained Elmo1. Clones were sent for Sanger sequencing, which confirmed the presence of ELMO1 in both vectors.

7.4.3 Making mutant *ELMO1* (F59L) by PCR

The next step was to make the F59L mutation in ELMO1. We generated the mutant ELMO1 (F59L) by site directed mutagenesis, in which we designed a primer that has the mutant codon TTA that codes for leucine (L) instead of the wt codon TTC that codes for the phenylalanine at position 59 (F59). PCR reactions (2.5 µl) were transformed into DH5α competent cells and plated on LB agar plates. Colonies were selected and plasmid DNA obtained using the Qiagen Mini-prep kit. DNA was sent for Sanger sequencing to confirm the presence of mutation (fig 7.9).

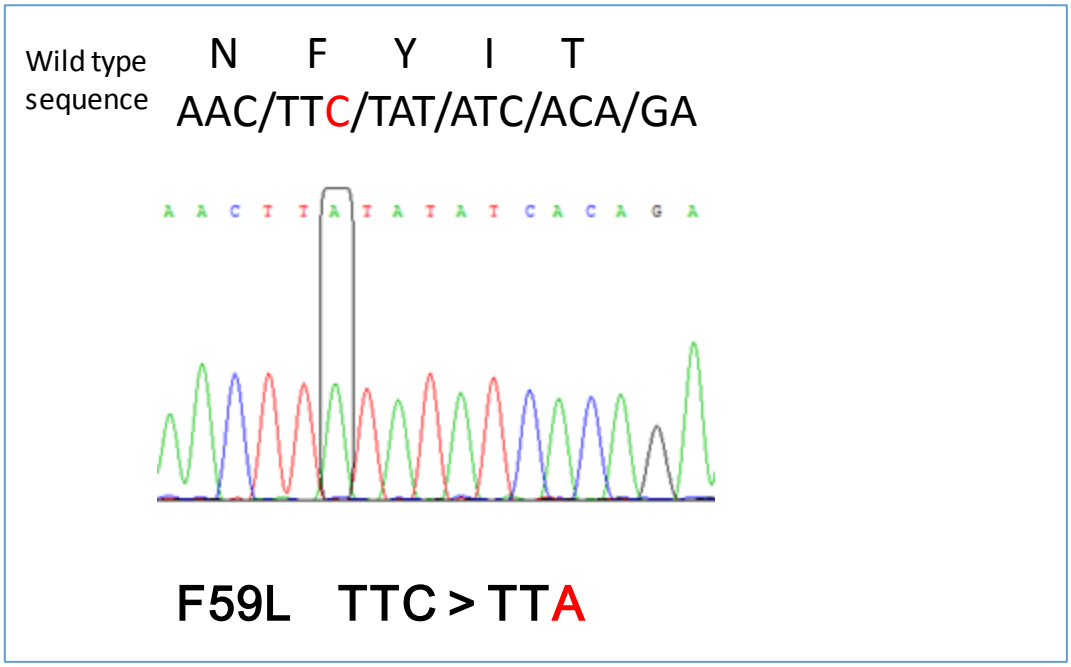


Figure 7.9. Confirmation of the presence of C>A mutation in ELMO1 by Sanger sequencing.

7.4.4 Transfection of wt and mutant *ELMO1* in OAC cell lines

We prepared large-scale plasmid DNA containing the wt and mutant *ELMO1* using the HiSpeed Maxi-prep kit from Qiagen. We transfected FLO-1 and OE19 cell lines with 1 µg of GFP empty vector, GFP-wt *ELMO1*, GFP-F59L *ELMO1*, Strep-wt *ELMO1*, or strep-F59L *ELMO1*, using Attractene as a tranfection reagent. After 24 hours of incubation at 37°C, cells were lysed with 0.5% NP40 lysis buffer. Protein concentrations were measured by Bradford assay, and 10µg of proteins were loaded on SDS-PAGE and immunoblotted with GFP and ELMO1 antibodies to test the expression of ELMO1. Wt and mutant ELMO1 were successfully expressed from both vectors in FLO-1 (fig 7.10) and OE19.

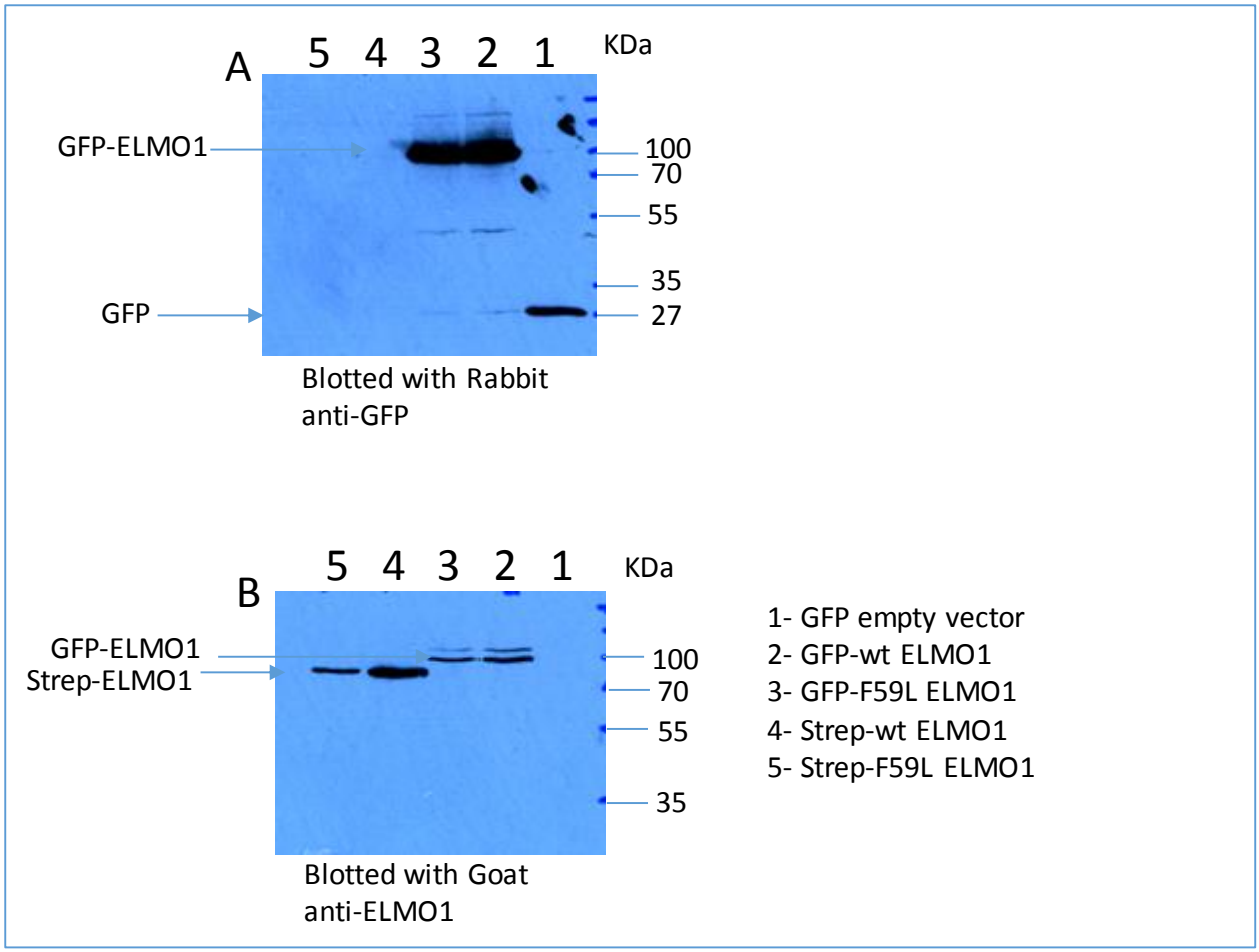


Figure 7.10. Western blot results of FLO-1 cell line transfected with wt and mutant ELMO1. *FLO-1 was transfected with 1µg of GFP empty vector (1) as a control, GFP-wt ELMO1 (2), GFP-F59L ELMO1 (3), Strep-wt ELMO1 (4), and Strep-F59L ELMO1 (5). Attractene was used as a transfection reagent. Cells were lysed 24h after incubation at 37°C, with 0.5% NP40 lysis buffer. Proteins (10µg) were run on two 12% gradient gels and then transferred to 0.2 µM nitrocellulose membrane. A, the first membrane was blotted with rabbit anti-GFP primary antibody, and swine anti-rabbit secondary antibody. B, the second membrane was blotted with goat anti-ELMO1 primary antibody and rabbit anti-goat secondary antibody. All antibodies were used at 1/1000 dilution and incubated for 1 hour at RT. Antibody signal was detected using enhanced chemiluminescence (ECL).*

7.4.5 Clonogenic assay

The clonogenic assay has been established for more than 50 years. It enables an assessment of the differences in reproductive viability (capacity of cells to produce progeny) between control untreated cells and cells that have undergone various treatments such as exposure to ionising radiation, various chemical compounds or, in some cases, genetic manipulation [142]. Here we have used this assay to study the effect of wt and mutant ELMO1 on the viability and growth of FLO-1 cells by transfecting them with empty vectors (GFP and Streptactin), vectors with wt ELMO1, and vectors with mutant ELMO1 (F59L). After 24h of transfection, cells were trypsinized and re-cultured in medium containing neomycin for growth selection; both the GFP and strep vectors contain neomycin resistant gene, so only cells that have these vectors will survive. After a period of time (around 1 month) cells were washed and dyed with Leishman stain. There were more colonies in the plates containing wt ELMO1, compared to those containing empty vector (for both GFP and strep). Mutant F59L ELMO1 results in a higher number of colonies (similar to the mutant P53) compared to the wt ELMO1 (fig 7.11). This means that the exogenous expression of wt ELMO1 has increased the growth of FLO-1 cells giving them advantageous growth compared with cells with no expression, and the F59L mutation increased the growth of cells even more than the wt ELMO1, meaning that it is a gain of function mutation. This is supported by another study, which shows that the silencing of ELMO1 or DOCK180 induced an appreciable inhibition of colony formation in SKOV3 compared with wild type, non-targeting, and control cells, which indicates a role for ELMO1 in cell growth (fig. 7.12 A, B) [131]. The same study investigated the roles of ELMO1 and DOCK180 on cell invasion by a Transwell Matrigel assay. As indicated in figure 6.12 C, D, ELMO1 or DOCK180 knockdown resulted in significant reduction of invasive cell number compared with control groups.

To further investigate the pathways of ELMO1 and its interacting proteins, we developed an SBP-tagged affinity purification method, in combination with label-free SWATH MS, to identify novel binding proteins of wt and mutant ELMO1 in the OAC cell lines FLO-1 and OE19.

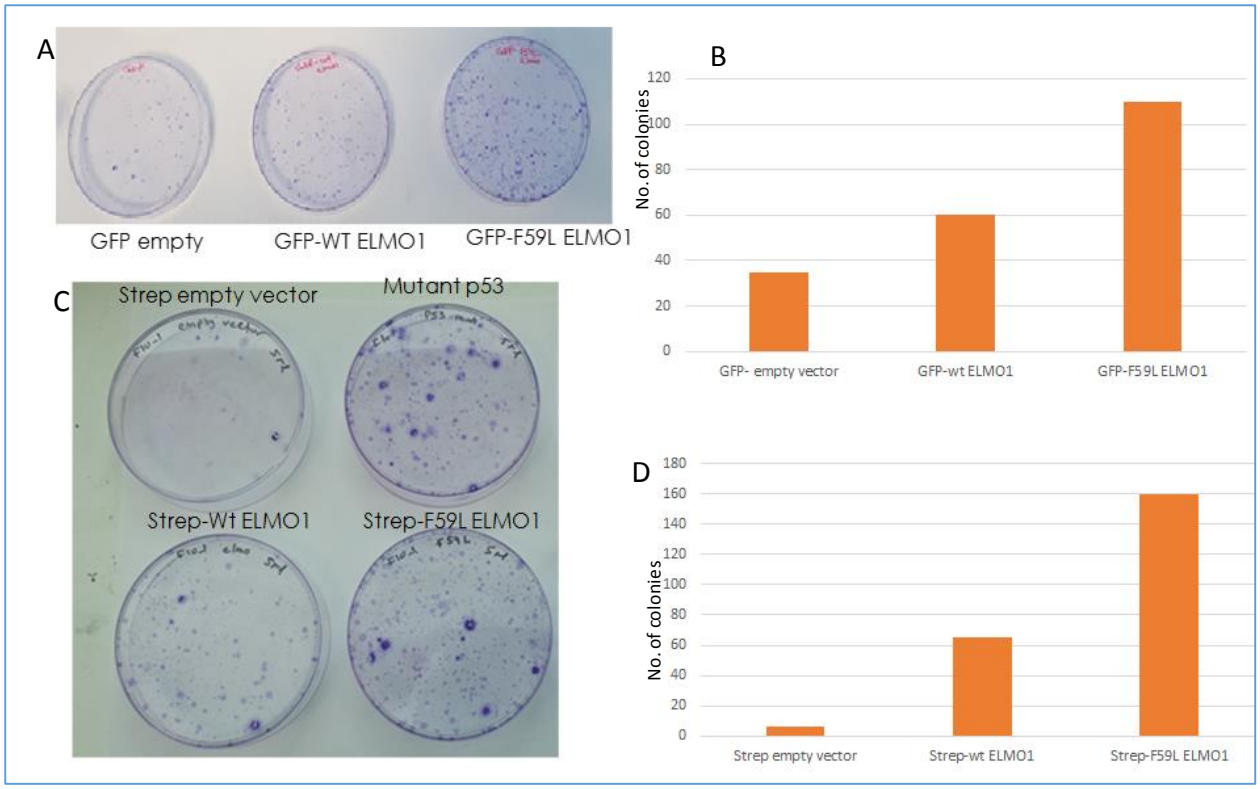


Figure 7.11. Clonogenic assay of FLO-1 cells transfected with control empty vectors and vectors with wt and mutant (F59L) ELMO1. A, B *FLO-1 cells transfected with GFP empty vector, GFP-wt ELMO1 or GFP-F59L ELMO1.* ***C, D*** *FLO-1 cells were transfected with Strep empty vector, Strep-wt ELMO1, Strep-F59L ELMO1 and Strep-mutant p53 gene (control). Cells were grown in media containing neomycin as selective drug. After one month of growth plates were dyed with Leishman stain and colonies were counted.*

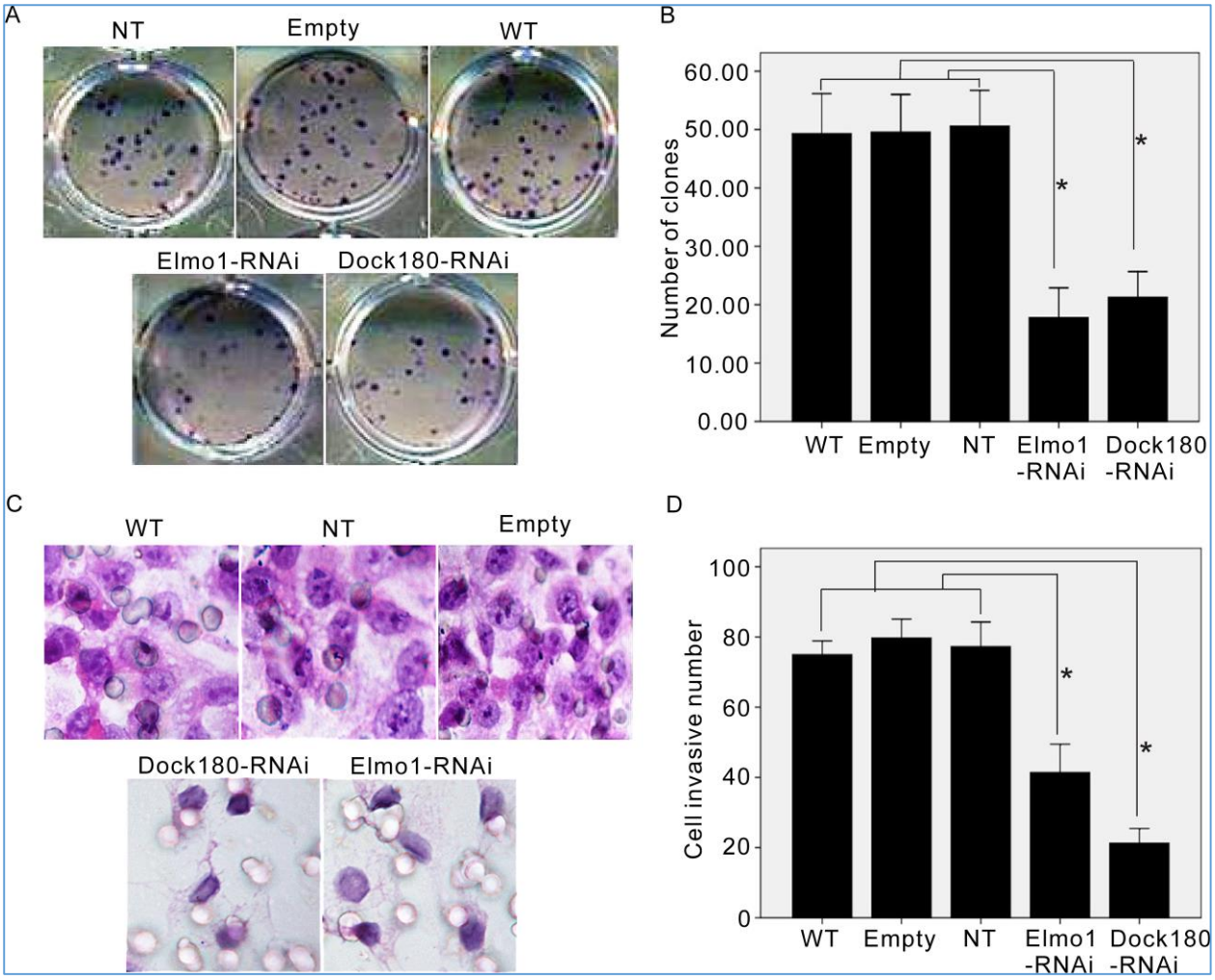


Figure 7.12. Cell proliferation and invasion in vitro in each cell group. ***A and B****, Dock180-RNAi and Elmo1-RNAi cells showed a significant reduction in their ability to form colonies as compared with NT control cells, empty vector control cells, and WT control cells. *P G 0.01.* ***C,*** *crystal violet stain of invading cells on the reverse side of the filter was shown (magnification, x 400). The bar analysis is shown in* ***(D).*** *Both cells silenced by Dock180 and Elmo1 had reduced the invasion of SKOV3 cells, compared with the control groups. *P G 0.05 [131].*

7.4.6 SBP-tagged pull down experiment

FLO-1 and OE19 cells were cultured in 10 cm plates and transfected with 5 µg of the Strep-tag vector, empty, wt *ELMO1* or F59L *ELMO1*, using Attractene transfection agent. After 24h of incubation at 37°C, cells were lysed with 500 µl of 0.5NP40 lysis buffer. Then, 15 µl of Streptavidin agarose beads were added to the lysate and incubated, shaking, at 4°C for 2 hours. The beads were then washed 3X with 0.5NP40 buffer and bound proteins were separated SDS–PAGE. Silver staining, and immunoblotting with ELMO1 specific antibodies were carried out; both detected the presence of ELMO1 on the beads that were incubated with lysates of cells transfected with Strep-tagged *ELMO1* (fig 7.13). In order to identify ELMO1 binding proteins, label-free SWATH-MS was performed, (in collaboration between the Hupp lab and Dr Borek Vojtesek's lab, Brno, Czech Republic) to identify the total proteins bound to the beads, which in turn bound to wt ELMO1 and F59L ELMO1.

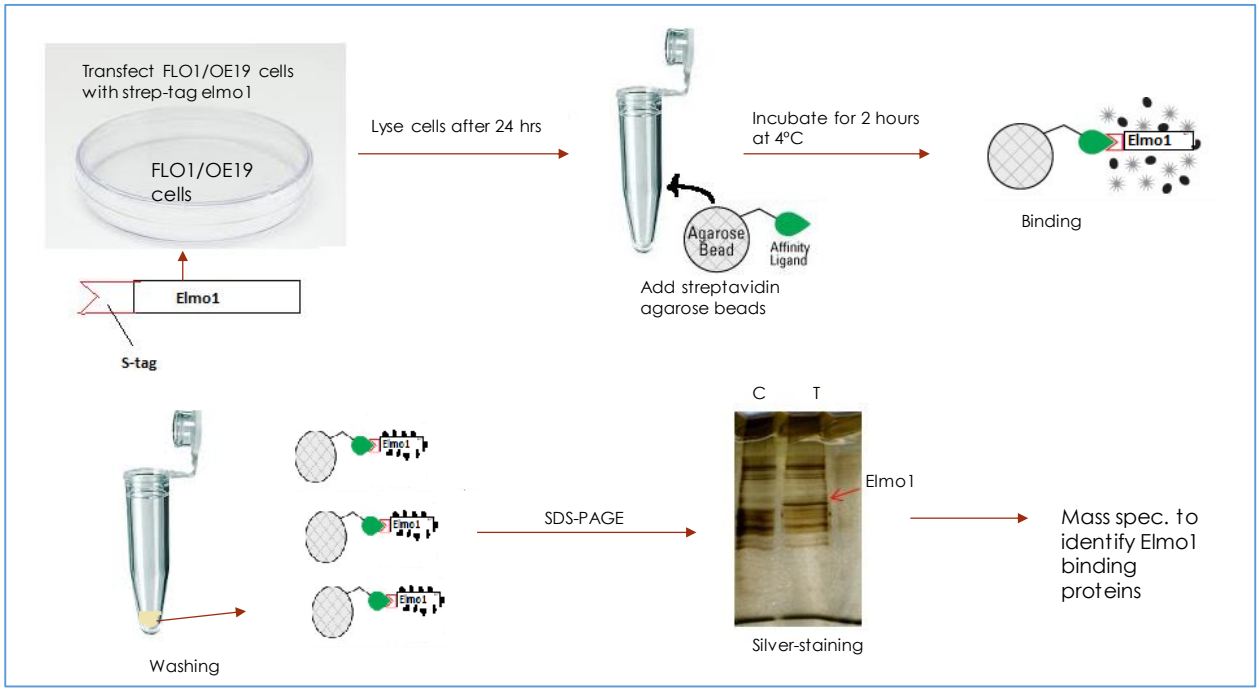


Figure 7.13. SBP-tagged pull down experiment with wt and mutant ELMO1. *FLO-1 and OE19 cells were transfected with Strep-tag vector (empty), Strep-tag wt ELMO1 or Strep-tag F59L ELMO1 and incubated for 24 hours at 37°C. The cells were then lysed with 500µl of 0.5% NP40 lysis buffer and 15µl of streptavidin agarose beads were added to the lysate and rotated for 2 hours at 4°C. The beads were collected and washed three times with NP40 buffer. SDS–PAGE lysis buffer was added to beads and run on 12% gradient gels. Silver staining and Western blot were performed to confirm the presence of ELMO1. The beads were sent for SWATH-MS to identify ELMO1 binding proteins.*

7.4.7 Detecting of ELMO1 binding proteins by SWATH-MS

The SWATH-MS method identified many proteins in the pull-down samples containing wt and mutant ELMO1 in both FLO-1 and OE19 cells. ELMO1 is detected in all samples, which confirms the analysis method, but with a lower level of detection in the mutant ELMO1 samples than in the wt ELMO1 samples (fig 7.14). The method identified previously known oncogenes as novel binding proteins of ELMO1, such as anterior gradient protein 2 (AGR2) and Dermcidin (DCD). When we compared the list of proteins detected in F59L ELMO1 to the list of the wt ELMO1, in both cell lines, it appears that the fold change of the binding of these two proteins is higher in mutant ELMO1 than in wt ELMO1 (fig 6.14). AGR2 has 2.93 of fold change in OE19 cells (table 7.1) and 3.07 in FLO-1. DCD has 6.94 in OE19 (table 7.1) and 3.29 in FLO-1. This means that the F59L mutation in ELMO1 increases its binding to other oncogene proteins, which explains the increase in the number of colonies and invasion of mutant ELMO1, compared with wild type protein.

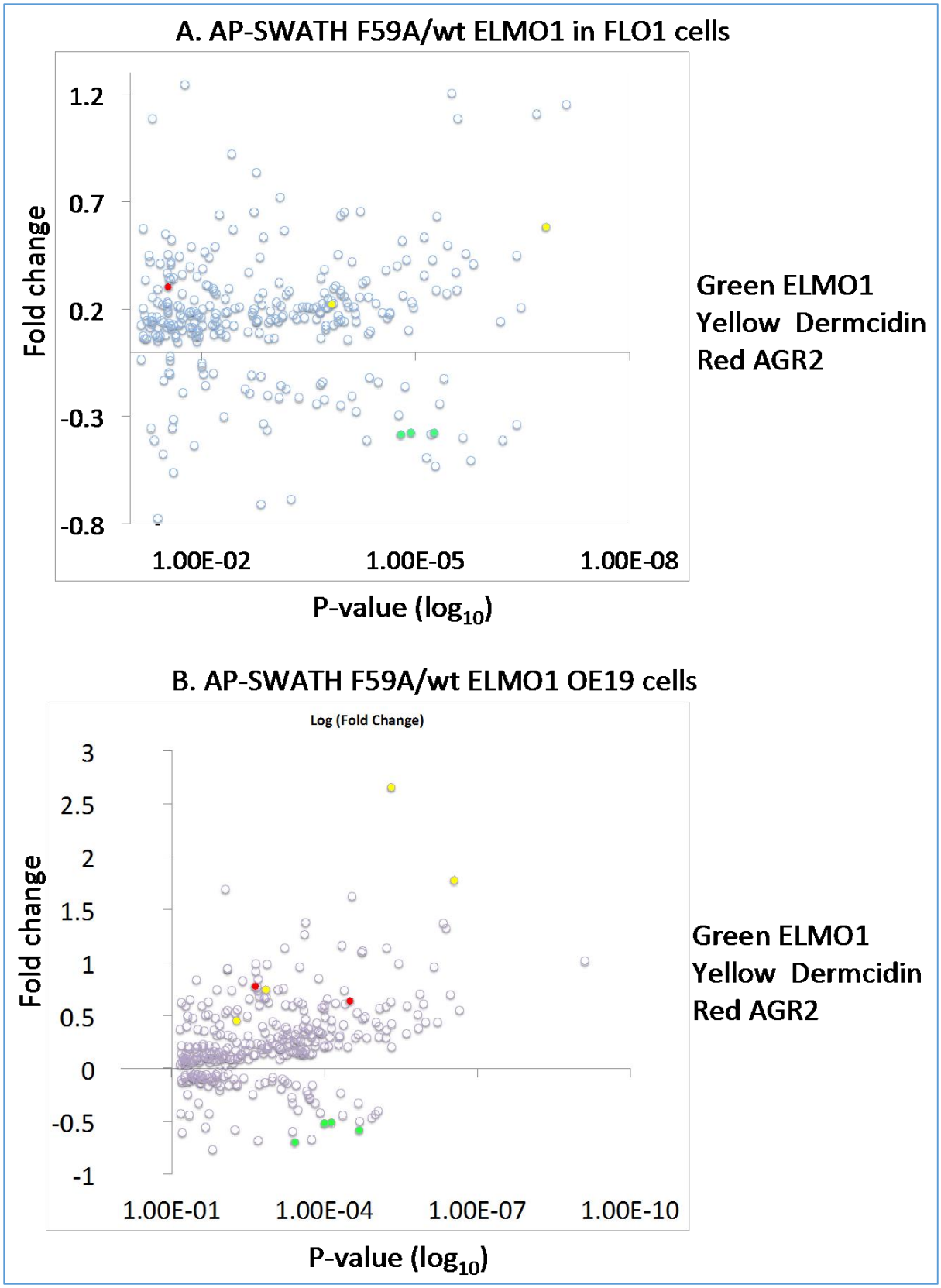


Figure 7.14. Scatter Blots of the fold change of proteins detected by SWATH-MS bound to the F59L ELMO1 over proteins detected with wt ELMO1 in FLO-1 (A) and OE19 (B). *ELMO1 (green) was detected with lower fold change in F59L ELMO1 in both cell lines. AGR2 (red) and Dermcidin (yellow) have higher fold change in the mutant ELMO1 than in the wt ELMO1 pull-downs from both cell lines.*

Number	Group	p-value	Fold Change
1	Plakophilin-1 GN=PKP1	0.00013	9.90911655020281
2	Ig gamma-1 chain C region GN=IGHG1	5.3021e-6	9.66379234318151
3	Cystatin-A GN=CSTA	0.00081	8.98575135827092
4	Dermcidin GN=DCD	7.6267e-6	6.9409570409027
5	60 KDa heat shock protein, mitochondrial OS=Homo sapiens GN=HSPD1 PE=1 SV=2	0.14099	6.72607337403442
6	Calmodulin-like protein 5 GN=CALML5	0.00321	6.54839654019489
7	Serum albumin GN=ALB	3.6991e-7	6.00598165464845
8	Aldehyde dehydrogenase, dimeric NADP-preferring GN=ALDH3A1	0.00089	5.86928863718585
9	Protein IGKV3-11 GN=IGKV3-11	1.9184e-5	4.98086211748401
10	3-ketoacyl-CoA thiolase GN=HADHB	0.00725	3.58790506078299
11	Filaggrin GN=FLG	0.00269	3.08113701964857
12	Isoform 3 of Tropomyosin beta chain GN=TPM2	0.01746	3.03133882700613
13	Eukaryotic translation initiation factor 4 gamma 1 GN=EIF4G1	0.08196	24.1044053307855
14	Corneodesmosin GN=CDSN	0.06600	2.96913023950455
15	Anterior gradient protein 2 homolog GN=AGR2	4.6897e-5	2.9319572452782
16	Isoform 2 of Mitochondrial import inner membrane translocase subunit TIM50 GN=TIMM50	0.20079	2.80697042919826
17	Zinc-alpha-2-glycoprotein GN=AZGP1	0.10022	2.79284140151035
18	Immunoglobulin lambda-like polypeptide 5 GN=IGLL5	0.00082	2.68818046681556
19	Arginase-1 GN=ARG1	0.00137	2.62664980897023
20	Keratin, type II cytoskeletal 78 GN=KRT78	0.00174	2.61137739849788
21	Fatty acid-binding protein, epidermal GN=FABP5	5.4751e-5	2.51704420641723
22	Junction plakoglobin GN=JUP	1.2009e-6	2.51530980971425
23	Prelamin-A/C GN=LMNA	0.03584	2.43842717447525
24	Tubulin beta-4B chain GN=TUBB4B	0.11432	2.40590529000604
25	Desmoplakin GN=DSP	6.3563e-5	2.28217367853831
26	Neuroblast differentiation-associated protein AHNAK GN=AHNAK	0.03380	2.24737343919498
27	Protein-glutamine gamma-glutamyltransferase E GN=TGM3	0.02862	2.20022119011577
28	Protein UBBP4 OS=Homo sapiens GN=UBBP4	0.00334	2.19133636029379
29	14-3-3 protein zeta/delta (Fragment) GN=YWHAZ	0.05568	2.13635213743807
30	Keratin, type II cytoskeletal 1b GN=KRT77	0.13516	2.11011803323294
31	Annexin A2 GN=ANXA2	0.01202	2.10609092114317
32	UPF0568 protein C14orf166 GN=C14orf166	0.12004	2.01366978621334

Table 7.1. Proteins that bound to the mutant ELMO1 (F59L) with >2-fold change compared to that in the wt ELMO1. *P-value and fold change mutant/wt ELMO1 are shown.*

7.4.8 Validation of the ELMO1–AGR2 interaction

We have carried out a proximity ligation assay (PLA) to validate the binding between ELMO1 and AGR2 proteins in FLO-1 cell lines. As mentioned earlier, FLO-1 does not express ELMO1, and we could not detect AGR2 expression in FLO-1 by Western blotting (AGR2 expression may be low in these cell lines). Therefore, for the PLA we had to transfect FLO-1 cells with both genes. We transfected FLO-1 cells, grown on cover slips in 6 well plates, with 1µg of the Strep-empty vector/Strep-ELMO1 (wt and mutant) with and without 1µg of pSF-CMV-AGR2 (provided by Ayman Mohd in Prof. Ted Hupp's group). PLA validated the binding between ELMO1 (both wt and mutant) and AGR2, as the fluorescent signals were only seen when both genes were expressed in FLO-1 cells (fig 7.15).

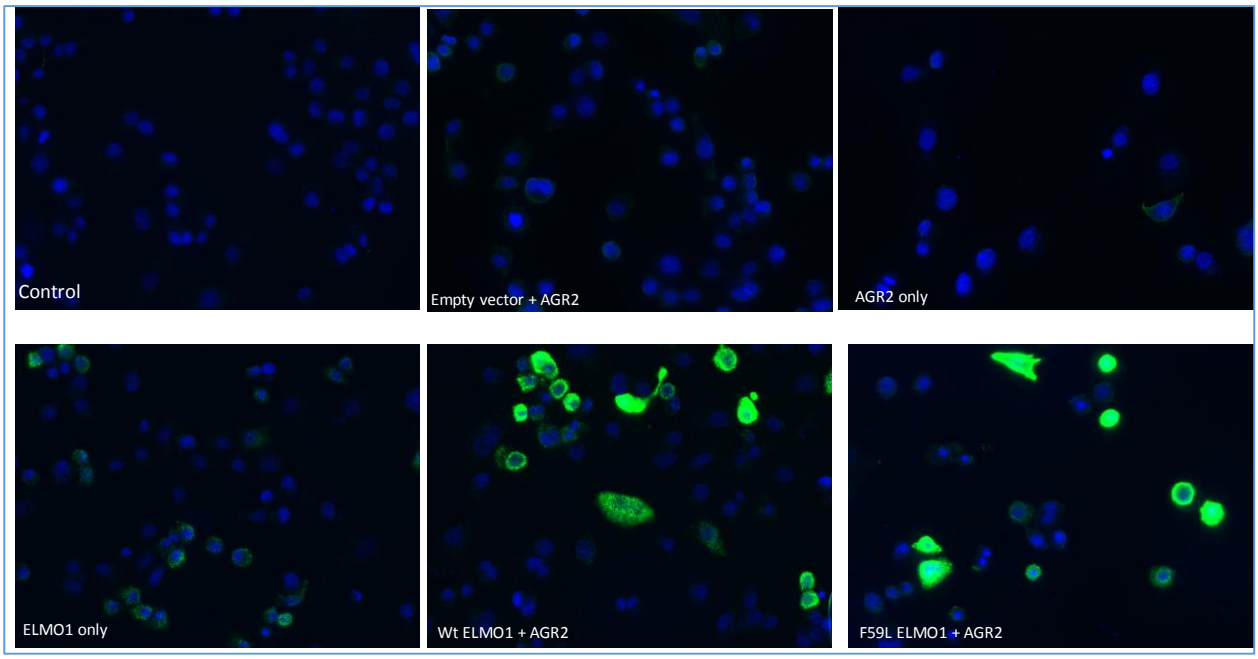


Figure 7.15. Proximity ligation assay (PLA) results in FLO-1 cells showing Duolink immunospots of ELMO1 (wt and mutant) with AGR2*. FLO-1 cells were grown on cover slips and transfected with control (no transfection), or 1µg of the following: Strep-empty vector + pSF-CMV-AGR2; pSF-CMV-AGR2 only; Strep-wt ELMO1 only; Strep-wt ELMO1 + pSF-CMV-AGR2; Strep-F59L ELMO1 + pSF-CMV-AGR2. They were blotted with goat polyclonal ELMO1 antibody + rabbit polyclonal AGR2 antibody.*

7.4.9 Co-expression of ELMO1 and AGR2

FLO-1 cells were transfected with 1 µg of strep-empty vector + AGR2, strep-wt ELMO1 + AGR2, strep-F59L ELMO1 + AGR2, and AGR2 only (pSF-CMV-AGR2) in 6-well plate, in order to see if there is any different in the expression of AGR2 when expressed with wt and mutant ELMO1. 24 hours after transfection, cells were lysed with urea lysis buffer. 15 µg of proteins were run on SDS-PAGE for Western Blot. The results of Western Blot show that there is no difference in the expression of AGR2 when expressed alone or with empty vector and when expressed with wt and mutant ELMO1 (fig 7.16).

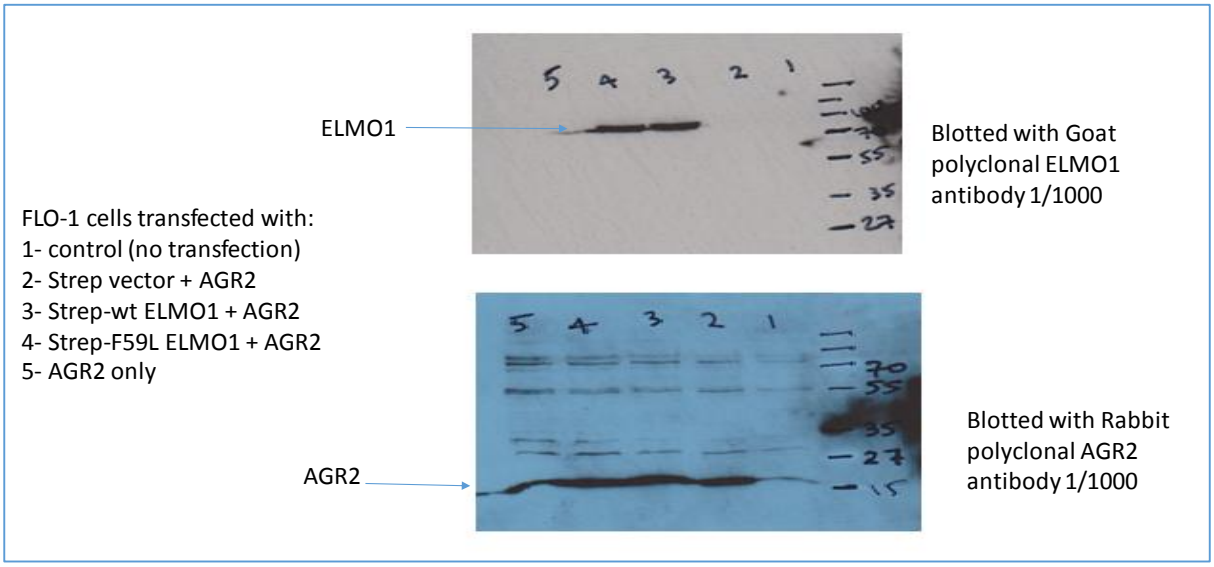


Figure 7.16. Western Blot results of co-expression of ELMO1 and AGR2 in FLO-1 cells*. FLO-1 cells were transfected with 1 µg of AGR2 with and without 1 µg of wt and mutant ELMO1 using attractine. After 24 h cells were lysed with urea lysis buffer and 15 µg of proteins of each of the control and the transfected cells were run on two SDS-PAGE. The first membrane was blotted with goat polyclonal ELMO1 primary antibody and rabbit anti goat secondary antibody. The second membrane was blotted with rabbit polyclonal AGR2 primary antibody and swine anti rabbit secondary antibody. Antibody signal was detected using enhanced chemiluminescence (ECL).*

7.4.10 OAC tissue microarray (TMA) of AGR2 and AGR2-induced FLO-1 tandem mass tag (TMT)

A proteomics study was carried out on two OAC cases to identify high expression proteins in the tumour, compared to the normal tissue. AGR2 was one of the elevated proteins in the tumour tissues of both cases (fig 7.17A), and this was then confirmed by tissue microarray (TMA), which confirmed the high expression of AGR2 in OAC tissues (fig 7.17B).

A further study was carried out to detect highly induced proteins in FLO-1 cells after the induction of AGR2. Tandem mass tag (TMT) was carried out on FLO-1 cells that were transfected with AGR2 or empty vector. The results showed that DCD was one of the proteins highly induced in the AGR2 positive cells but not in the negative control (data not shown), which suggests that AGR2 induces the expression of DCD.

Both of these studies were performed in collaboration between the Hupp lab and Dr Borek Vojtesek's lab, Brno, Czech Republic.

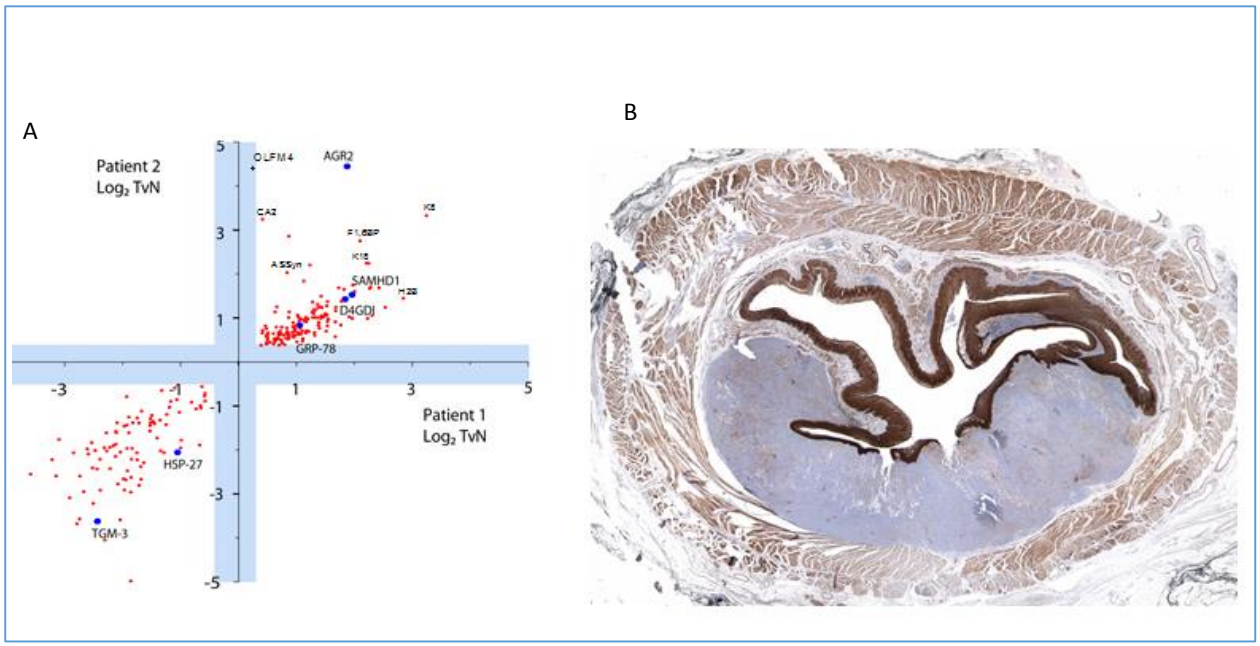


***Figure 7.17. High expression of AGR2 in the tumour tissues of OAC patients. A**, AGR2 is detected among the highly expressed proteins in tumour tissues compared to the normal tissues of two OAC patients. **B**, TMA result of OAC tissue using AGR2 antibody*

7.5 Discussion

7.5.1 ELMO1 interaction with AGR2 and DCD

Here we report, for the first time, that ELMO1 protein binds to the oncogenic proteins AGR2 and DCD, and their binding is greater to the mutant form of ELMO1 (F59L). It is still not known if this binding to ELMO1 is direct or via other protein(s). The F59L mutation is in the N-terminal of ELMO1 where active RhoG binds to ELMO1 and relieves ELMO1 from autoinhibition and transfers the ELMO1–DOCK complex from the cytoplasm to the cell membrane, where it activates RAC1. This mutation leads to more binding with AGR2 and DCD and shows increased cell growth and invasion, compared with wt ELMO1. Binding of AGR2 and DCD may help ELMO1 to increase binding to RhoG, stabilizing the complex to the cell membrane (fig 7.18), and thus increasing ELMO1 activity. The mutation may change the structure of ELMO1 protein in a way that makes it more exposed to AGR2 and DCD.

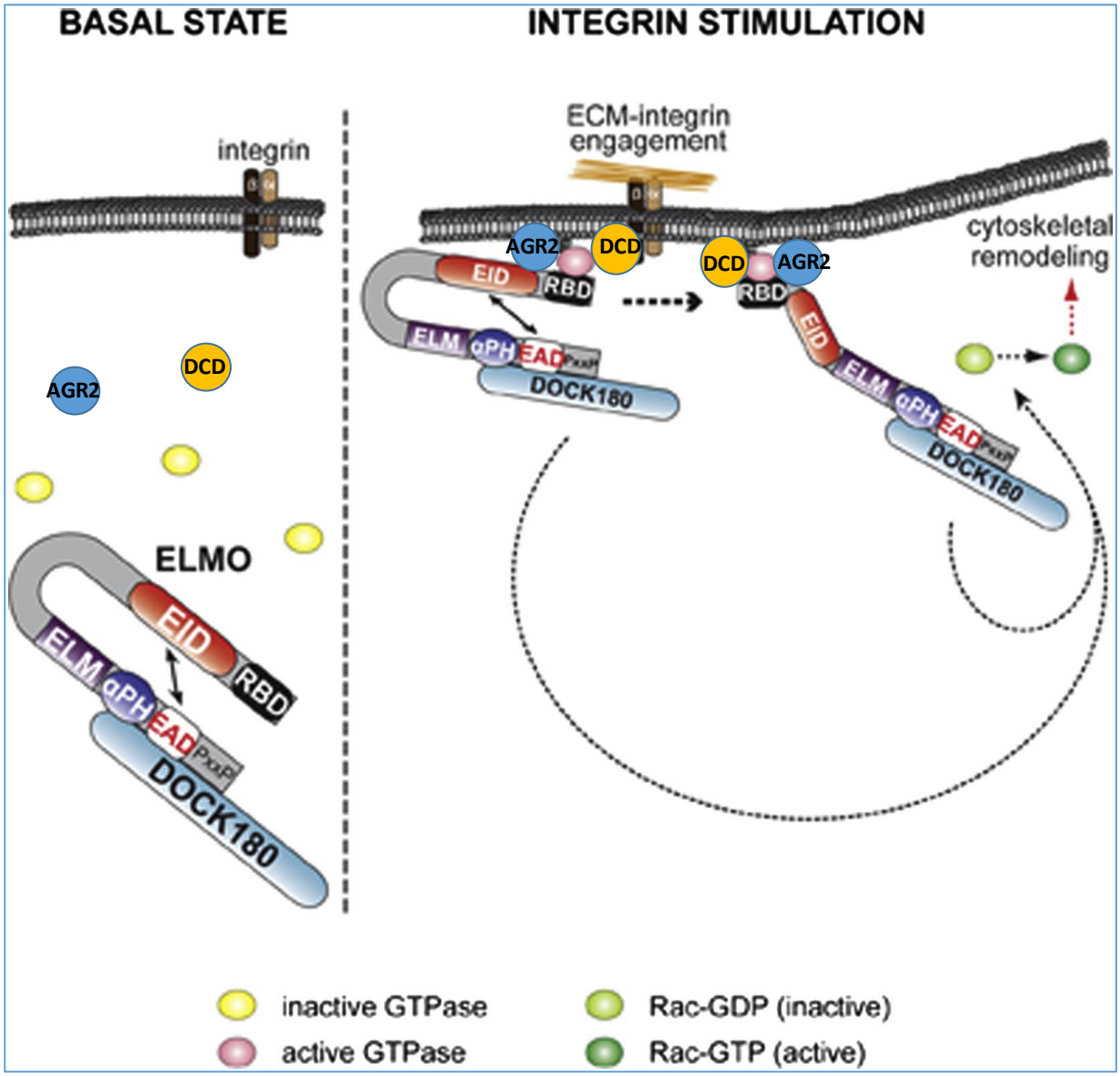


Figure 7.18. Model of ELMO1 binding to AGR2 and DCD*. AGR2 and DCD may help to stabilize the ELMO1–DOCK complex to bind to the activated RhoG and activate RAC1.*

7.5.2 AGR2 production and function

AGR2 is a 17 KDa protein that is highly conserved in vertebrates. In humans, enhanced AGR2 expression was first described in breast cancer, which was followed by similar observations in most human adenocarcinomas, including those derived from oesophagus, pancreas, lung, ovary, and prostate. Both in vitro and in vivo studies demonstrated that AGR2 promotes tumour growth and metastasis [143]. AGR2 is classified as a member of the protein disulphide isomerase (PDI) family of proteins on the basis of amino acid sequence homology. It is localised to the endoplasmic reticulum (ER) as well as secreted. It has also been found in the nucleus and on the cell surface [144]. PDI enzymes contain CXXC domains, which function in

oxidation/reduction reactions and isomerisation of disulphide bridges, thereby facilitating the maturation of proteins in the ER and ensuring correct folding and multimerisation of proteins targeted for the secretory pathway [144, 145]. AGR2 can be found outside cells, such as in the blood or the urine of cancer patients, which led some researchers to suggest that measuring the levels of AGR2 in body fluids may be a useful marker for detecting cancers [146]. It was reported that AGR2 makes cancer cells more aggressive, specifically when found outside cells [146]. Secretion of AGR2 was reported to correlate with metastasis and poor prognosis in breast cancer, and considered as a biomarker in prostate cancer. AGR2 upregulation was also detected in pancreatic carcinoma tissues. In gastric cancer tissues, higher expression in gastric cancer cells has previously been reported to be evident in the cytoplasm compared with non-tumour cells. Notably, AGR2 can be used as a suitable candidate gene for the detection of circulating tumour cells, a novel resource for the identification of molecular markers in many types of cancer patients [147].

7.5.3 AGR2 cellular mechanisms

The cellular mechanism by which AGR2 promotes growth are poorly understood, but there have been a number of observations that are beginning to shed light on the role it plays in cell signalling networks. One pathway that AGR2 has been implicated in is the EGFR pathway. EGFR is a well-characterized receptor-tyrosine kinase that functions in development and serves a vital role in many human cancers. It was found that binding of EGFR to AGR2 in the ER is required for receptor delivery to the plasma membrane and thus EGFR signalling. AGR2 was reported to induce expression of AREG, a growth promoting EGFR ligand. This study functionally linked AGR2 and AREG and supported a significant role for AGR2 in lung adenocarcinomas and the regulation of cell growth. As a result of AGR2 expression, AREG may stimulate the EGFR signalling pathway and may be responsible for the increased cell proliferation and anchorage-independent growth observed in transformed cells [148, 149].

AGR2 exhibits the basic features of a pro-oncogenic protein (fig 7.19). It was reported that AGR2 is upstream of, and stimulates, key cancer-signalling pathways, such as cyclin D1, c-Myc, p-Src and survivin, based on using both short interfering RNA and ectopic expression to manipulate AGR2 levels [148]. Genomic analysis of AGR2-stable cells by cDNA microarray revealed that AGR2 overexpression upregulates the expression of genes involved in cell

proliferation, invasion and angiogenesis, which are very important for tumour progression and metastasis [148].

7.5.4 AGR2 and p53 silencing

In OAC there is an early selection pressure for p53 mutation during the replacement of squamous epithelium with metaplastic epithelium, which is called 'Barrett's oesophagus'. To identify potentially novel p53 inhibitors in this cancer 'intermediate', a clinical proteomics screen had been set up in a proliferative disease (Barrett's epithelium) and identified AGR2, which was validated as a potent inhibitor of p53-dependent transcription and a growth-promoting proto-oncogene, which suggests that AGR2 pathway might be a key mechanism to silence p53 signalling [148].

Both increased ELMO1 and AGR2 expression has shown an increase in the growth and metastasis of many cancer cell lines. The interaction between the two proteins (either direct or via another protein) may lead to further increase in tumour behaviour and metastasis. The effect of their interaction on cell growth and migration need to be further studied as their interaction region may serve as a therapeutic target in OAC and other cancers.

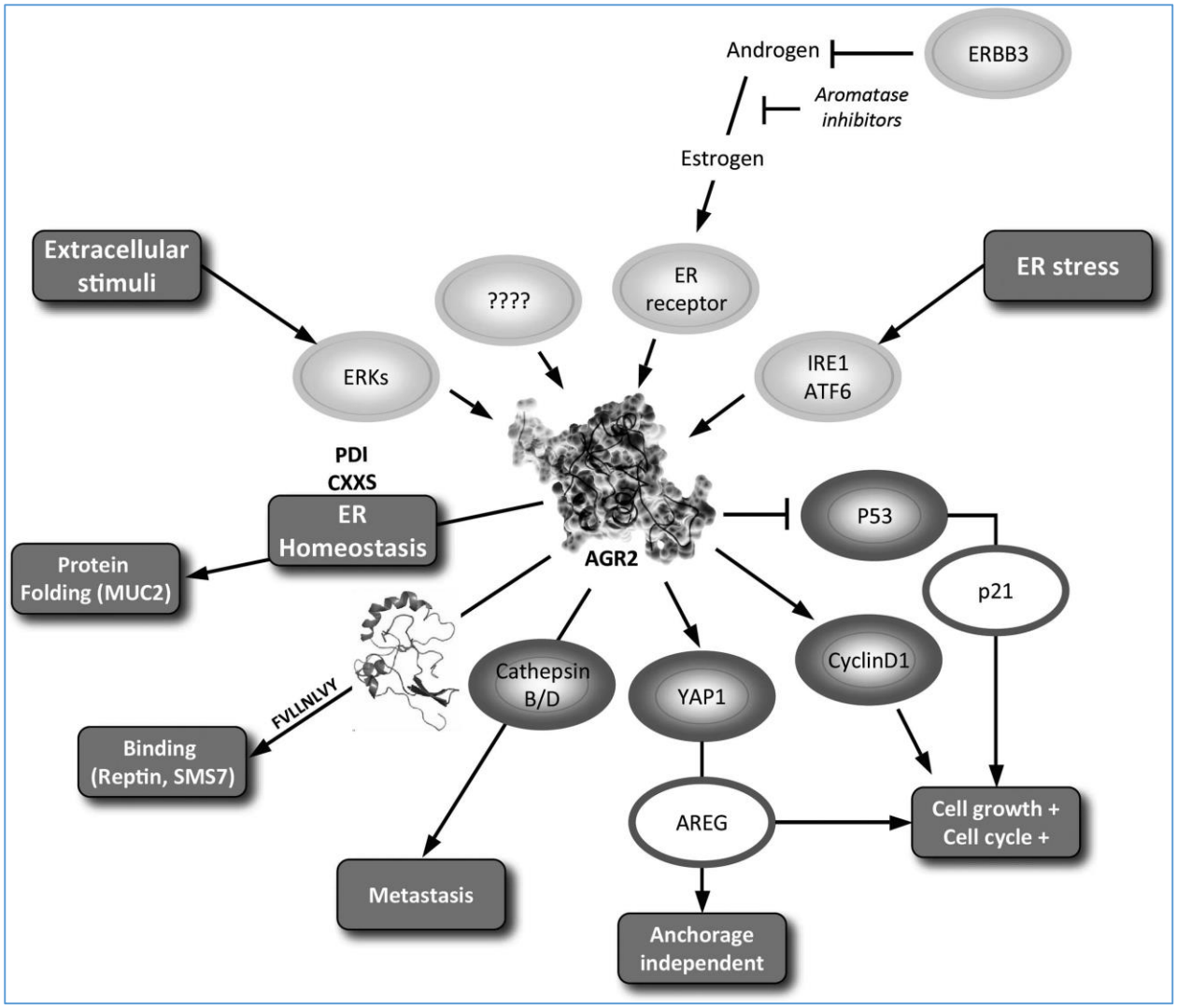


Figure 7.19. AGR2 biological pathways. *AGR2 is presented as the central point of this picture (structure derived from ERP1831). Pathway intermediates known to regulate AGR2 expression are indicated in light grey ovals. AGR2-dependent functions are indicated in dark grey ovals. Potential (not experimentally demonstrated) intermediates are indicated as empty ovals. Physiological/pathophysiological inputs or outputs are represented as grey boxes.*

7.5.5 DCD production and function

DCD is another oncogene that is found to bind ELMO1 in OAC cell lines. The DCD gene encodes for an 11 KDa protein, which in normal tissues displays a restricted expression pattern, with significant expression detected only in eccrine sweat glands of the skin, and in certain parts of the brain [150]. Overexpression of DCD was reported in multiple cancers such as melanoma, breast, prostate, pancreatic and hepatocellular carcinoma (HCC) [150]. DCD has diverse biological functions, such as acting as a growth and survival factor in breast cancer and in neural cells, displaying antibacterial activity, and inducing cancer associated cachexia in animal models and in cancer patients [150].

7.5.6 Binding of DCD and ELMO1 to Nck1 in hepatocellular carcinoma tissues

A study on HCC has identified DCD as a novel binding protein of Nck1 by performing GST pull-down assays using the SH2 domain of Nck1 as bait. The Nck family adaptor proteins couple tyrosine phosphorylation signals to actin cytoskeletal reorganization that leads to cell motility [151]. This study has reported that the binding of DCD to the SH2 domain of Nck1 was only detected in the tumour tissues and moreover that DCD expression was restricted to tumour tissues. They reported that the overexpression of wild type DCD resulted in a remarkable increase in the activation of both RAC1 and CDC42 compared with control cells, which supports a role of DCD in tumour metastasis. Interestingly, when we examined the list of the other proteins (table 7.2) that were detected to bind to the Nck1, we found that ELMO1 is one of these proteins and also that it is only detected in the tumour tissues, similar to DCD. It was reported that phosphotyrosine residues at position 18, 216, 395, and 511 of ELMO1 mediate binding to the SH2 domain of NcK1. The association of Nck1 with ELMO1 facilitated the binding of ELMO1 to active RhoG, and thus promoted the GEF activity of ELMO1–DOCK180 complex [152]. So, both ELMO1 and DCD bind the same domain of Nck1, or alternatively these two proteins may bind to each other and one of them binds to Nck1. We do not have Nck1 in the SWATH-MS list of ELMO1 binding proteins in OAC cells, so DCD may bind to ELMO1 even in the absence of Nck1. This binding may be direct or through a third protein such as AGR2.

Number	SwissProt ID	Protein name	No. of matched peptide	T/N ratio
1	Q9NTJ3	Structural maintenance of chromosomes protein 4	4	>2
2	P16152	Carbonyl reductase [NADPH] 1	4	>2
3	Q86YZ3	Hornerin	5	>2
4	P60660	Myosin light polypeptide 6	1	>2
5	Q969J3	Loss of heterozygosity 12 chromosomal region 1	3	>2
6	P62736	Actin, aortic smooth muscle	2	>2
7	Q99943	1-acyl-sn-glycerol-3-phosphate acyltransferase alpha	1	nd N
8	Q6T4R5	Nance-Horan syndrome protein	2	nd N
9	Q86YZ3	Hornerin	12	nd N
10	Q99583	Max-binding protein MNT	2	nd N
11	P60709	Actin, cytoplasmic 1	10	nd N
12	P81605	Dermcidin	2	nd N
13	Q92556	Engulfment and cell motility protein 1	1	nd N
14	P60059	Protein transport protein Sec61 subunit gamma	4	nd N

Table 7.2. Differentially expressed Nck-SH2 binding proteins in HCC tissues. *nd N, not detected in normal tissue [151]*

7.6 Conclusion

Methods were devised to begin to rapidly and quantitatively discover new functions of “orphan” mutated proteins in cancer. Steps in the method include: (i) clonogenic assay; (ii) SBP-pull down; and (iii) proximity ligation assays to validate the interaction in vivo. I suggest that many of the mutated proteins I discuss in my thesis, such MAP4K5, could also be processed using this workflow. In this case study, mutated ELMO was chosen as the model. Exogenous expression of ELMO1 was shown to increase the number of colonies of the OAC cell lines OE19 and FLO-1. This number was increased more with expression of the F59L mutation of ELMO1 that was detected by whole exome sequencing of OAC tissues, suggesting this is a gain-of-function mutation. The SWATH-MS of pull-downs of SBP-tagged wt and mutant ELMO1 has shown that ELMO1 binds to the oncogenic proteins AGR2 and DCD, and this binding is greater to the mutant form of ELMO1. This interaction between ELMO1 and the two proteins AGR2 and DCD needs to be further studied to investigate its effects on proliferation and migration of cancer cells, and it may work as a therapeutic target in OAC and other cancers.

CHAPTER EIGHT
CONCLUSION AND FUTURE WORK

8.1 Detection of mutations using CLC-bio software

We have used a novel software algorithm, CLC-bio to analyse the cancer genome by DNAseq and expressed cancer genome arising from transcription by RNAseq to define dominant sources of potentially expressed tumour-specific mutations and oncogenic targets. We focused primarily on the rare human pleomorphic sarcoma as a disease of high unmet clinical need but used a range of cancer models to accelerate the development of the pipeline.

8.1.1 Analysing the whole exomes sequences of patients with head and neck tumour

First, we applied next generation sequencing of whole exomes of tumour tissues and two matched normal tissues (blood and "normal" tumour adjacent tissue) from a small set of patients with head and neck cancer to define parameters for use of the new software. The approaches identified significant mutations in tumour relative to germline DNA, but also in normal adjacent tissue, relative to normal germline, consistent with known field cancerization. The analysis results showed that there were two different genetic groups in the five patients, independent of the age of the patients. The first group has mutations in the p53 and have more CNVs, which is seen in the old patients 119 and 137, and in the young patient 82. The other group does not have mutations in p53 and has very few CNVs.

8.1.2 Analysing the whole exomes of twenty cancer pleomorphic sarcoma cancer patients

For setting up the larger sequencing screen in the subsequent set of twenty cancer pleomorphic sarcoma cancer patients, whole exome sequencing was performed on tumour tissues and their matched normal adjacent tissues, rather than germline blood derived DNA, to define truly tumour-specific mutations. This approach provided sets of recurrent non-synonymous mutations in tumour tissue such as a transmembrane protease and suggests potential therapeutic targets for future focus that are highly tumour specific in pleomorphic sarcoma.

8.1.3 CLC-Bio and MuTect

We have shown that the CLC-bio software has detected higher number of variants in each sarcoma patients compared to MuTect software which might be due to the different parameters applied by the two softweres. These detected variants need to be further validated by Sanger sequences to make sure that they are real somatic variants.

8.1.4 Cancer heterogeneity

Two biopsies taken from different regions of one sarcoma tumour have shown different somatic mutations when compared to the normal adjacent tissue of the same patient, which is an evidence of intratumour heterogeneity. However, the two different lists of genes share several pathways such as RAS pathway and Wnt signalling pathway.

8.1.5 Expressed somatic mutations in the RNA

Identifying the expressed cancer genome, using RNAseq and protein quantitation methods such as mass spectrometry, provides a more accurate view of the state of the cancer tissue at the time of presentation in the clinic. The CLC-Bio software has enabled us to identify the dominant RNAs that were cancer genome encoded, and led us to highlight ~15% of mutated cancer genes are expressed at the time of surgery. I suggest that these mutated pathways or genes represent more realistic targets for therapeutics or diagnostics than conventional mutated cancer genes. For example, a mutant gene might have been required very early in the evolution of the cancer and might not even be expressed at the time of surgery. In addition to the expressed cancer-encoded genome, we also were able to use the novel software to identify mutations in RNA that are not genome encoded. This is suggestive of RNA editing. Some of these genes were validated by Sanger sequencing and by examining RNAseq in cell lines. For example, we found that MAP4K5 is indeed edited in cell lines, thus providing a model gene to dissect stages in RNA editing in the future.

8.1.6 expression of somatic mutations at the protein level

Currently, we do not know how many of the cancer genes that are mutated, the expressed RNA that is genome encoded or the RNA that is edited, are expressed at the protein level. I have shown one example where the mutated PDCD6 gene had detectable mutated RNA and mutated peptide using mass spectrometry. The full exploitation of mass spectrometry was beyond the scope of my PhD thesis, which was focused largely on DNA and RNA sequencing. It will nevertheless be interesting in future to determine how much of the potential mutation, whether coded from genome mutation or RNA editing, is translated into mutated protein. My current thesis data shows, using novel software designed for biologists and clinicians, that is it realistic to routinely define this expressed genome at the nucleic acid level to support future proteomics studies.

8.1.7 Conclusion

In conclusion, we have applied and validated newly emerging software to begin to interrogate cancer tissue from patients of unmet clinical need in order to define new mechanisms of cancer progression and to define possibly new or better drug targets for new therapies. The data identified highly recurrent genome encoded mutations in human pleomorphic sarcoma and a potentially novel, targetable landscape represented by RNA editing driven mutant protein production. This will provide a foundation for future work on making better choices to advance our ability to improve patient management in human pleomorphic sarcoma.

8.2 Applied methods to study the effects of wt and mutant ELMO1

Lastly, novel or orphan mutant proteins observed in human cancers, whether from DNA encoded mutant proteins or from RNA-edited driven mutant protein synthesis require new tools and technologies to discover new oncogenic signalling mechanisms. Methods were devised to begin to rapidly and quantitatively discover new functions of “orphan” mutated proteins in cancer. Steps in the method include: (i) clonogenic assay; (ii) SBP-pull down; and (iii) proximity ligation assays to validate the interaction in vivo. I suggest that many of the mutated proteins I discussed in my thesis, such MAP4K5 and PDCD6, could also be processed using this workflow. In this case study, mutated ELMO was chosen as the model. Exogenous expression of ELMO1 was shown to increase the number of colonies of the OAC cell lines OE19 and FLO-1. This number was increased more with expression of the F59L mutation of ELMO1 that was detected by whole exome sequencing of OAC tissues, suggesting this is a gain-of-

function mutation. The SWATH-MS of pull-downs of SBP-tagged wt and mutant ELMO1 has shown that ELMO1 binds to the oncogenic proteins AGR2 and DCD, and this binding is greater to the mutant form of ELMO1. This interaction between ELMO1 and the two proteins AGR2 and DCD needs to be further studied to investigate its effects on proliferation and migration of cancer cells, and it may work as a therapeutic target in OAC and other cancers.

8.3 Summary for future work

- Study the effects of the detected somatic mutations in key genes such as PDCD6 on cells growth and invasion by the methods used in the study; clonogenic assay
- From the mass-spectrometry data; detect which mutated genes have higher expression in the tumour compared to the normal adjacent tissues, and look if the mutation has been translated into the peptide
- Detection of the list of proteins that bind to the mutated gene by using the SWATH-MS of pull-downs of SBP-tagged of the wild type and mutant forms of the gene
- Validation the RNA editing events, and study their effects on splicing and on cells growth and number of colony formation
- Perform tumour microarray of ELMO1 and DCD in OAC tissues to detect the expression of ELMO1 and DCD and find out if there is an overlap between the expression of ELMO1, DCD and AGR2
- Study the effects of binding of ELMO1 to AGR2 and DCD on cells proliferation and migration
- Detection of the binding regions of ELMO1 to the AGR2 and DCD

CHAPTER NINE
References

1. Sudhakar, A., *History of Cancer, Ancient and Modern Treatment Methods.* J Cancer Sci Ther, 2009. **1**(2): p. 1-4.
2. Cornelisse, C.J. and P. Devilee, *Facts in cancer genetics.* Patient Educ Couns, 1997. **32**(1-2): p. 9-17.
3. Torre, L.A., et al., *Global cancer statistics, 2012.* CA Cancer J Clin, 2015. **65**(2): p. 87-108.
4. Stewart, B.W., et al., *Cancer prevention as part of precision medicine: 'plenty to be done'.* Carcinogenesis, 2016. **37**(1): p. 2-9.
5. Talseth-Palmer, B.A. and R.J. Scott, *Genetic variation and its role in malignancy.* Int J Biomed Sci, 2011. **7**(3): p. 158-71.
6. Stratton, M.R., *Exploring the genomes of cancer cells: progress and promise.* Science, 2011. **331**(6024): p. 1553-8.
7. Thomas, D.M., et al., *Cancer 2015: a longitudinal whole-of-system study of genomic cancer medicine.* Drug Discov Today, 2015. **20**(12): p. 1429-32.
8. Harrington, K.J., *The biology of cancer.* Medicine, 2016. **44**(1): p. 1-5.
9. Podlaha, O., et al., *Evolution of the cancer genome.* Trends Genet, 2012. **28**(4): p. 155-63.
10. LeBlanc, V.G. and M.A. Marra, *Next-Generation Sequencing Approaches in Cancer: Where Have They Brought Us and Where Will They Take Us?* Cancers (Basel), 2015. **7**(3): p. 1925-58.
11. Rizzo, J.M. and M.J. Buck, *Key principles and clinical applications of "next-generation" DNA sequencing.* Cancer Prev Res (Phila), 2012. **5**(7): p. 887-900.
12. Beltran, H., et al., *Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response.* JAMA Oncol, 2015. **1**(4): p. 466-74.
13. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends Genet, 2008. **24**(3): p. 133-41.
14. Reuter, J.A., D.V. Spacek, and M.P. Snyder, *High-throughput sequencing technologies.* Mol Cell, 2015. **58**(4): p. 586-97.
15. Chaitankar, V., et al., *Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research.* Prog Retin Eye Res, 2016.
16. Lohmann, K. and C. Klein, *Next generation sequencing and the future of genetic diagnosis.* Neurotherapeutics, 2014. **11**(4): p. 699-707.
17. Wilkerson, M.D., et al., *Integrated RNA and DNA sequencing improves mutation detection in low purity tumors.* Nucleic Acids Res, 2014. **42**(13): p. e107.
18. Han, S.S., et al., *RNA sequencing identifies novel markers of non-small cell lung cancer.* Lung Cancer, 2014. **84**(3): p. 229-35.
19. Kukurba, K.R. and S.B. Montgomery, *RNA Sequencing and Analysis.* Cold Spring Harb Protoc, 2015. **2015**(11): p. 951-69.
20. Whitley, S.K., W.T. Horne, and J.K. Kolls, *Research Techniques Made Simple: Methodology and Clinical Applications of RNA Sequencing.* J Invest Dermatol, 2016. **136**(8): p. e77-82.
21. Lee, E.Y. and W.J. Muller, *Oncogenes and tumor suppressor genes.* Cold Spring Harb Perspect Biol, 2010. **2**(10): p. a003236.
22. Zhu, K., et al., *Oncogenes and tumor suppressor genes: comparative genomics and network perspectives.* BMC Genomics, 2015. **16 Suppl 7**: p. S8.
23. Josephidou, M., A.G. Lynch, and S. Tavare, *multiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples.* Nucleic Acids Res, 2015. **43**(9): p. e61.

24. Martincorena, I. and P.J. Campbell, *Somatic mutation in cancer and normal cells.* Science, 2015. **349**(6255): p. 1483-9.
25. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
26. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
27. Alexandrov, L.B. and M.R. Stratton, *Mutational signatures: the patterns of somatic mutations hidden in cancer genomes.* Curr Opin Genet Dev, 2014. **24**: p. 52-60.
28. Segovia, R., A.S. Tam, and P.C. Stirling, *Dissecting genetic and environmental mutation signatures with model organisms.* Trends Genet, 2015. **31**(8): p. 465-74.
29. Farkona, S., E.P. Diamandis, and I.M. Blasutig, *Cancer immunotherapy: the beginning of the end of cancer?* BMC Med, 2016. **14**: p. 73.
30. Schumacher, T.N. and R.D. Schreiber, *Neoantigens in cancer immunotherapy.* Science, 2015. **348**(6230): p. 69-74.
31. Emens, L.A., et al., *Cancer immunotherapy trials: leading a paradigm shift in drug development.* J Immunother Cancer, 2016. **4**: p. 42.
32. Kvistborg, P., et al., *Immune monitoring technology primer: whole exome sequencing for neoantigen discovery and precision oncology.* Journal for ImmunoTherapy of Cancer, 2016. **4**(1).
33. Hackl, H., et al., *Computational genomics tools for dissecting tumour-immune cell interactions.* Nat Rev Genet, 2016. **17**(8): p. 441-58.
34. Schumacher, T.N. and N. Hacohen, *Neoantigens encoded in the cancer genome.* Curr Opin Immunol, 2016. **41**: p. 98-103.
35. Tureci, O., et al., *Targeting the Heterogeneity of Cancer with Individualized Neoepitope Vaccines.* Clin Cancer Res, 2016. **22**(8): p. 1885-96.
36. Dodd, R.D., *Emerging targets in sarcoma: Rising to the challenge of RAS signaling in undifferentiated pleomorphic sarcoma.* Cancer, 2016. **122**(1): p. 17-9.
37. Contino, G., et al., *Whole-genome sequencing of nine esophageal adenocarcinoma cell lines.* F1000Res, 2016. **5**: p. 1336.
38. Mansour, N.M., S.S. Groth, and S. Anandasabapathy, *Esophageal Adenocarcinoma: Screening, Surveillance, and Management.* Annual review of medicine, 2016.
39. Qian, J. and J.-Y. Fang, *Genetic Variations in Esophageal Cancer.* Gastrointestinal Tumors, 2015. **2**(3): p. 124-130.
40. Garcia, E., et al., *Authentication and characterisation of a new oesophageal adenocarcinoma cell line: MFD-1.* Sci Rep, 2016. **6**: p. 32417.
41. Dulak, A.M., et al., *Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity.* Nat Genet, 2013. **45**(5): p. 478-86.
42. Weaver, J.M., et al., *Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis.* Nat Genet, 2014. **46**(8): p. 837-43.
43. Gillet, L.C., et al., *Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis.* Mol Cell Proteomics, 2012. **11**(6): p. O111 016717.
44. Kang, H., A. Kiess, and C.H. Chung, *Emerging biomarkers in head and neck cancer in the era of genomics.* Nat Rev Clin Oncol, 2015. **12**(1): p. 11-26.
45. Toporcov, T.N., et al., *Risk factors for head and neck cancer in young adults: a pooled analysis in the INHANCE consortium.* Int J Epidemiol, 2015. **44**(1): p. 169-85.
46. Krishnan, N., et al., *Integrated analysis of oral tongue squamous cell carcinoma identifies key variants and pathways linked to risk habits, HPV, clinical parameters and tumor recurrence.* F1000Res, 2015. **4**: p. 1215.
47. Agrawal, N., et al., *Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1.* Science, 2011. **333**(6046): p. 1154-7.

48. Riaz, N., et al., *Unraveling the molecular genetics of head and neck cancer through genome-wide approaches.* Genes Dis, 2014. **1**(1): p. 75-86.
49. Lokker, M.E., et al., *Symptoms of patients with incurable head and neck cancer: prevalence and impact on daily functioning.* Head Neck, 2013. **35**(6): p. 868-76.
50. Brand, T.M., et al., *AXL Is a Logical Molecular Target in Head and Neck Squamous Cell Carcinoma.* Clin Cancer Res, 2015. **21**(11): p. 2601-12.
51. Stransky, N., et al., *The mutational landscape of head and neck squamous cell carcinoma.* Science, 2011. **333**(6046): p. 1157-60.
52. Zhang, S., J.S. Wei, and J. Khan, *The Significance of Transcriptome Sequencing in Personalized Cancer Medicine.* 2014: p. 49-64.
53. Castle, J.C., et al., *Mutated tumor alleles are expressed according to their DNA frequency.* Sci Rep, 2014. **4**: p. 4743.
54. Shah, S.P., et al., *The clonal and mutational evolution spectrum of primary triple-negative breast cancers.* Nature, 2012. **486**(7403): p. 395-9.
55. Braakhuis, B.J., et al., *A genetic explanation of Slaughter's concept of field cancerization: evidence and clinical implications.* Cancer Res, 2003. **63**(8): p. 1727-30.
56. Furuya, T., et al., *CNVs Associated with Susceptibility to Cancers: A Mini-Review.* Journal of Cancer Therapy, 2015. **06**(05): p. 413-422.
57. Shlien, A. and D. Malkin, *Copy number variations and cancer.* Genome Med, 2009. **1**(6): p. 62.
58. Willis, J.A., et al., *Genome-wide analysis of the role of copy-number variation in pancreatic cancer risk.* Front Genet, 2014. **5**: p. 29.
59. Pickering, C.R., et al., *Squamous cell carcinoma of the oral tongue in young non-smokers is genomically similar to tumors in older smokers.* Clin Cancer Res, 2014. **20**(14): p. 3842-8.
60. Li, R., et al., *Clinical, genomic, and metagenomic characterization of oral tongue squamous cell carcinoma in patients who do not smoke.* Head Neck, 2015. **37**(11): p. 1642-9.
61. Hayes, D.N., C. Van Waes, and T.Y. Seiwert, *Genetic Landscape of Human Papillomavirus-Associated Head and Neck Cancer and Comparison to Tobacco-Related Tumors.* J Clin Oncol, 2015. **33**(29): p. 3227-34.
62. Vettore, A.L., et al., *Mutational landscapes of tongue carcinoma reveal recurrent mutations in genes of therapeutic and prognostic relevance.* Genome Med, 2015. **7**(1): p. 98.
63. Ow, T.J., et al., *Effective Biomarkers and Radiation Treatment in Head and Neck Cancer.* Arch Pathol Lab Med, 2015. **139**(11): p. 1379-88.
64. Suh, Y., et al., *Clinical update on cancer: molecular oncology of head and neck cancer.* Cell Death Dis, 2014. **5**: p. e1018.
65. Alhejaily, A., et al., *Inactivation of the CDKN2A tumor-suppressor gene by deletion or methylation is common at diagnosis in follicular lymphoma and associated with poor clinical outcome.* Clin Cancer Res, 2014. **20**(6): p. 1676-86.
66. McIlwain, D.R., T. Berger, and T.W. Mak, *Caspase functions in cell death and disease.* Cold Spring Harb Perspect Biol, 2013. **5**(4): p. a008656.
67. Kuwahara, D., et al., *Caspase-9 regulates cisplatin-induced apoptosis in human head and neck squamous cell carcinoma cells.* Cancer Lett, 2000. **148**(1): p. 65-71.
68. Park, J.M., et al., *MSH3 mismatch repair protein regulates sensitivity to cytotoxic drugs and a histone deacetylase inhibitor in human colon carcinoma cells.* PLoS One, 2013. **8**(5): p. e65369.
69. Paliwal, S., et al., *CtBP2 Promotes Human Cancer Cell Migration by Transcriptional Activation of Tiam1.* Genes Cancer, 2012. **3**(7-8): p. 481-90.
70. Bergman, L.M., et al., *Role of the unique N-terminal domain of CtBP2 in determining the subcellular localisation of CtBP family proteins.* BMC Cell Biol, 2006. **7**: p. 35.
71. Byun, J.S. and K. Gardner, *C-Terminal Binding Protein: A Molecular Link between Metabolic Imbalance and Epigenetic Regulation in Breast Cancer.* Int J Cell Biol, 2013. **2013**: p. 647975.

72. Takayama, K., et al., *CtBP2 modulates the androgen receptor to promote prostate cancer progression.* Cancer Res, 2014. **74**(22): p. 6542-53.
73. Mirnezami, A.H., et al., *Hdm2 Recruits a Hypoxia-Sensitive Corepressor to Negatively Regulate p53-Dependent Transcription.* Current Biology, 2003. **13**(14): p. 1234-1239.
74. Rothenberg, S.M. and L.W. Ellisen, *The molecular pathogenesis of head and neck squamous cell carcinoma.* Journal of Clinical Investigation, 2012. **122**(6): p. 1951-1957.
75. Andersson, C., et al., *Profiling of potential driver mutations in sarcomas by targeted next generation sequencing.* Cancer Genet, 2016. **209**(4): p. 154-60.
76. Mohan, M. and N. Jagannathan, *Oral field cancerization: an update on current concepts.* Oncol Rev, 2014. **8**(1): p. 244.
77. Dakubo, G.D., et al., *Clinical implications and utility of field cancerization.* Cancer Cell Int, 2007. **7**: p. 2.
78. Murphy, K.L., A.P. Dennis, and J.M. Rosen, *A gain of function p53 mutant promotes both genomic instability and cell survival in a novel p53-null mammary epithelial cell model.* FASEB J, 2000. **14**(14): p. 2291-302.
79. Taylor, B.S., et al., *Advances in sarcoma genomics and new therapeutic targets.* Nat Rev Cancer, 2011. **11**(8): p. 541-57.
80. Silveira, S.M., et al., *Genomic signatures predict poor outcome in undifferentiated pleomorphic sarcomas and leiomyosarcomas.* PLoS One, 2013. **8**(6): p. e67643.
81. Guijarro, M.V., et al., *Dual Pten/Tp53 suppression promotes sarcoma progression by activating Notch signaling.* Am J Pathol, 2013. **182**(6): p. 2015-27.
82. Barretina, J., et al., *Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy.* Nat Genet, 2010. **42**(8): p. 715-21.
83. Grimer, R., et al., *Guidelines for the management of soft tissue sarcomas.* Sarcoma, 2010. **2010**: p. 506182.
84. Desar, I.M., et al., *Advanced soft-tissue sarcoma and treatment options: critical appraisal of trabectedin.* Cancer Manag Res, 2016. **8**: p. 95-104.
85. Wardelmann, E., et al., *Soft tissue sarcoma: from molecular diagnosis to selection of treatment. Pathological diagnosis of soft tissue sarcoma amid molecular biology and targeted therapies.* Ann Oncol, 2010. **21 Suppl 7**: p. vii265-9.
86. Ballinger, M.L., et al., *Monogenic and polygenic determinants of sarcoma risk: an international genetic study.* The Lancet Oncology, 2016. **17**(9): p. 1261-1271.
87. Hofvander, J., et al., *Recurrent PRDM10 gene fusions in undifferentiated pleomorphic sarcoma.* Clin Cancer Res, 2015. **21**(4): p. 864-9.
88. Christoforides, A., et al., *Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs.* BMC Genomics, 2013. **14**: p. 302.
89. Alioto, T.S., et al., *A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing.* Nat Commun, 2015. **6**: p. 10001.
90. Madsen, D.H., et al., *TMPRSS13 deficiency impairs stratum corneum formation and epidermal barrier acquisition.* Biochem J, 2014. **461**(3): p. 487-95.
91. de Aberasturi, A.L. and A. Calvo, *TMPRSS4: an emerging potential therapeutic target in cancer.* Br J Cancer, 2015. **112**(1): p. 4-8.
92. Jin, Z., Y.X. Han, and X.R. Han, *The role of APOBEC3B in chondrosarcoma.* Oncol Rep, 2014. **32**(5): p. 1867-72.
93. Yan, S., et al., *Increased APOBEC3B Predicts Worse Outcomes in Lung Cancer: A Comprehensive Retrospective Study.* J Cancer, 2016. **7**(6): p. 618-25.
94. Burns, M.B., et al., *APOBEC3B is an enzymatic source of mutation in breast cancer.* Nature, 2013. **494**(7437): p. 366-70.
95. Fisher, R., L. Pusztai, and C. Swanton, *Cancer heterogeneity: implications for targeted therapeutics.* Br J Cancer, 2013. **108**(3): p. 479-85.
96. Gay, L., A.M. Baker, and T.A. Graham, *Tumour Cell Heterogeneity.* F1000Res, 2016. **5**.

97. Seoane, J. and L. De Mattos-Arruda, *The challenge of intratumour heterogeneity in precision medicine.* J Intern Med, 2014. **276**(1): p. 41-51.
98. McGranahan, N. and C. Swanton, *Biological and therapeutic impact of intratumor heterogeneity in cancer evolution.* Cancer Cell, 2015. **27**(1): p. 15-26.
99. Cibulskis, K., et al., *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotechnol, 2013. **31**(3): p. 213-9.
100. Wang, Q., et al., *Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers.* Genome Med, 2013. **5**(10): p. 91.
101. Finotello, F. and B. Di Camillo, *Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis.* Brief Funct Genomics, 2015. **14**(2): p. 130-42.
102. Wan, Q., et al., *BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis.* Database (Oxford), 2015. **2015**.
103. Conesa, A., et al., *A survey of best practices for RNA-seq data analysis.* Genome Biol, 2016. **17**: p. 13.
104. Shlien, A., et al., *Direct Transcriptional Consequences of Somatic Mutation in Breast Cancer.* Cell Rep, 2016. **16**(7): p. 2032-46.
105. Halabi, N.M., et al., *Preferential Allele Expression Analysis Identifies Shared Germline and Somatic Driver Genes in Advanced Ovarian Cancer.* PLoS Genet, 2016. **12**(1): p. e1005755.
106. Han, Y., et al., *Advanced Applications of RNA Sequencing and Challenges.* Bioinform Biol Insights, 2015. **9**(Suppl 1): p. 29-46.
107. Cirulli, E.T., et al., *Screening the human exome: a comparison of whole genome and whole transcriptome sequencing.* Genome Biol, 2010. **11**(5): p. R57.
108. Andries, V., et al., *NBPF1, a tumor suppressor candidate in neuroblastoma, exerts growth inhibitory effects by inducing a G1 cell cycle arrest.* BMC Cancer, 2015. **15**: p. 391.
109. Katz, C., et al., *Molecular basis of the interaction between proapoptotic truncated BID (tBID) protein and mitochondrial carrier homologue 2 (MTCH2) protein: key players in mitochondrial death pathway.* J Biol Chem, 2012. **287**(18): p. 15016-23.
110. Aureli, A., et al., *HLA-DRB1*13:01 allele in the genetic susceptibility to colorectal carcinoma.* Int J Cancer, 2015. **136**(10): p. 2464-8.
111. Guo, H., et al., *Association between polymorphisms in cdc27 and breast cancer in a Chinese population.* Tumour Biol, 2015. **36**(7): p. 5299-304.
112. Katoh, M., *Function and cancer genomics of FAT family genes (review).* Int J Oncol, 2012. **41**(6): p. 1913-8.
113. Ruggles, K.V., et al., *An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer.* Mol Cell Proteomics, 2016. **15**(3): p. 1060-71.
114. Ellis, M.J., et al., *Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium.* Cancer Discov, 2013. **3**(10): p. 1108-12.
115. Alfaro, J.A., et al., *Onco-proteogenomics: cancer proteomics joins forces with genomics.* Nat Methods, 2014. **11**(11): p. 1107-13.
116. Murray, E., et al., *Quantitative proteomic profiling of pleomorphic human sarcoma identifies CLIC1 as a dominant pro-oncogenic receptor expressed in diverse sarcoma types.* J Proteome Res, 2014. **13**(5): p. 2543-59.
117. Li, Z., et al., *Periostin expression and its prognostic value for colorectal cancer.* Int J Mol Sci, 2015. **16**(6): p. 12108-18.
118. Choudhury, A., et al., *Silencing of ROR1 and FMOD with siRNA results in apoptosis of CLL cells.* Br J Haematol, 2010. **151**(4): p. 327-35.
119. Huang, W.H., et al., *RNA editing and drug discovery for cancer therapy.* ScientificWorldJournal, 2013. **2013**: p. 804505.
120. Licht, K. and M.F. Jantsch, *Rapid and dynamic transcriptome regulation by RNA editing and RNA modifications.* J Cell Biol, 2016. **213**(1): p. 15-22.

121. Galeano, F., et al., *Human BLCAP transcript: new editing events in normal and cancerous tissues.* Int J Cancer, 2010. **127**(1): p. 127-37.
122. Avesson, L. and G. Barry, *The emerging role of RNA and DNA editing in cancer.* Biochim Biophys Acta, 2014. **1845**(2): p. 308-16.
123. Han, L. and H. Liang, *RNA editing in cancer: Mechanistic, prognostic, and therapeutic implications.* Mol Cell Oncol, 2016. **3**(2): p. e1117702.
124. Wang, O.H., et al., *Prognostic and Functional Significance of MAP4K5 in Pancreatic Cancer.* PLoS One, 2016. **11**(3): p. e0152300.
125. Yao, J., et al., *Overexpression of BLCAP induces S phase arrest and apoptosis independent of p53 and NF-kappaB in human tongue carcinoma : BLCAP overexpression induces S phase arrest and apoptosis.* Mol Cell Biochem, 2007. **297**(1-2): p. 81-92.
126. Hu, X., et al., *RNA over-editing of BLCAP contributes to hepatocarcinogenesis identified by whole-genome and transcriptome sequencing.* Cancer Lett, 2015. **357**(2): p. 510-9.
127. Mansour, N.M., S.S. Groth, and S. Anandasabapathy, *Esophageal Adenocarcinoma: Screening, Surveillance, and Management.* Annu Rev Med, 2016.
128. Hanawa-Suetsugu, K., et al., *Structural basis for mutual relief of the Rac guanine nucleotide exchange factor DOCK2 and its partner ELMO1 from their autoinhibited forms.* Proc Natl Acad Sci U S A, 2012. **109**(9): p. 3305-10.
129. Epting, D., et al., *The Rac1 regulator ELMO1 controls vascular morphogenesis in zebrafish.* Circ Res, 2010. **107**(1): p. 45-55.
130. Stevenson, C., et al., *Essential role of Elmo1 in Dock2-dependent lymphocyte migration.* J Immunol, 2014. **192**(12): p. 6062-70.
131. Wang, J., et al., *Elmo1 helps dock180 to regulate Rac1 activity and cell migration of ovarian cancer.* Int J Gynecol Cancer, 2014. **24**(5): p. 844-50.
132. Grimsley, C.M., et al., *Dock180 and ELMO1 proteins cooperate to promote evolutionarily conserved Rac-dependent cell migration.* J Biol Chem, 2004. **279**(7): p. 6087-97.
133. Bid, H.K., et al., *RAC1: an emerging therapeutic option for targeting cancer angiogenesis and metastasis.* Mol Cancer Ther, 2013. **12**(10): p. 1925-34.
134. Jarzynka, M.J., et al., *ELMO1 and Dock180, a bipartite Rac1 guanine nucleotide exchange factor, promote human glioma cell invasion.* Cancer Res, 2007. **67**(15): p. 7203-11.
135. Lee, J., et al., *Identification of a novel protein interaction between Elmo1 and Cdc27.* Biochem Biophys Res Commun, 2016. **471**(4): p. 497-502.
136. Patel, M., et al., *An evolutionarily conserved autoinhibitory molecular switch in ELMO proteins regulates Rac signaling.* Curr Biol, 2010. **20**(22): p. 2021-7.
137. Katoh, H. and M. Negishi, *RhoG activates Rac1 by direct interaction with the Dock180-binding protein Elmo.* Nature, 2003. **424**(6947): p. 461-4.
138. Lee, J., et al., *Arhgef16, a novel Elmo1 binding partner, promotes clearance of apoptotic cells via RhoG-dependent Rac1 activation.* Biochim Biophys Acta, 2014. **1843**(11): p. 2438-47.
139. Huang, X. and H. Townley, *Knock-down of ELMO1 in Paediatric Rhabdomyosarcoma Cells by Nanoparticle Mediated siRNA Delivery.* Nanobiomedicine, 2016: p. 1.
140. Li, H., et al., *Association between Galphai2 and ELMO1/Dock180 connects chemokine signalling with Rac activation and metastasis.* Nat Commun, 2013. **4**: p. 1706.
141. Janardhan, A., et al., *HIV-1 Nef binds the DOCK2-ELMO1 complex to activate rac and inhibit lymphocyte chemotaxis.* PLoS Biol, 2004. **2**(1): p. E6.
142. Rafehi, H., et al., *Clonogenic assay: adherent cells.* J Vis Exp, 2011(49).
143. Dong, A., et al., *The human adenocarcinoma-associated gene, AGR2, induces expression of amphiregulin through Hippo pathway co-activator YAP1 activation.* J Biol Chem, 2011. **286**(20): p. 18301-10.
144. Alavi, M., et al., *High expression of AGR2 in lung cancer is predictive of poor survival.* BMC Cancer, 2015. **15**: p. 655.

145. Dumartin, L., et al., *AGR2 is a novel surface antigen that promotes the dissemination of pancreatic cancer cells through regulation of cathepsins B and D.* Cancer Res, 2011. **71**(22): p. 7091-102.
146. Fessart, D., et al., *Secretion of protein disulphide isomerase AGR2 confers tumorigenic properties.* Elife, 2016. **5**.
147. Zhang, J., et al., *AGR2 is associated with gastric cancer progression and poor survival.* Oncol Lett, 2016. **11**(3): p. 2075-2083.
148. Chevet, E., et al., *Emerging roles for the pro-oncogenic anterior gradient-2 in cancer development.* Oncogene, 2013. **32**(20): p. 2499-509.
149. Dong, A., D. Wodziak, and A.W. Lowe, *Epidermal growth factor receptor (EGFR) signaling requires a specific endoplasmic reticulum thioredoxin for the post-translational control of receptor presentation to the cell surface.* J Biol Chem, 2015. **290**(13): p. 8016-27.
150. Bancovik, J., et al., *Dermcidin exerts its oncogenic effects in breast cancer via modulation of ERBB signaling.* BMC Cancer, 2015. **15**: p. 70.
151. Shen, S.L., et al., *Identification of Dermcidin as a novel binding protein of Nck1 and characterization of its role in promoting cell migration.* Biochim Biophys Acta, 2011. **1812**(6): p. 703-10.
152. Zhang, G., et al., *A novel interaction between the SH2 domain of signaling adaptor protein Nck-1 and the upstream regulator of the Rho family GTPase Rac1 engulfment and cell motility 1 (ELMO1) promotes Rac1 activation and cell motility.* J Biol Chem, 2014. **289**(33): p. 23112-22.